Impact de la manipulation thermique embryonnaire sur le méthylome de caille japonaise

Coralie Gimonnet, Anais Vitorino Carvalho, Nathalie Couroussé, Sabine Crochet, Thierry Bordeau, Marjorie Mersch, Benoit Piegu, Christelle Hennequet-Antier, Aurélien Brionne, Frederique Pitel, et al.

> Nantes
2—5
juillet

JOBIM 2019

JOURNÉES OUVERTES DE BIOLOGIE INFORMATIQUE & MATHÉMATIQUES

Thèmatiques :

Biologie structurale
Biologie des systèmes
Epidémiologie Génétique
Evolution/Phylogénie
Génomique/Métagénomique
Sciences des données

Abstracts

Keynotes :

Chloé-Agathe Azencott, Paris
Alexander Bockmayr, Berlin
Alessandra Carbone, Paris
Olivier Delaneau, Lausanne
Christophe Dessimoz, Lausanne
Juliette Martin, Lyon

https://jobim2019.sciencesconf.org

Dear attendees of the 20th edition of JOBIM, welcome in Nantes !

JOBIM is the French national conference dedicated to promoting an active interface between Biology, Computer Sciences, and Mathematics. After a previous visit to Nantes in 2009, JOBIM comes back this year in this same city from western France. Since the last visit, the bioinformatics community has impressively grown, and new fields are today covered. The impressive number of submissions at JOBIM 2019 reflects such an increase in our community. The Program Committee received a total of 260 submissions deciphered as 29 long presentations, 11 flash presentations, 15 demos and 204 posters. As a main novelty of the 2019th edition, JOBIM 2019 will present five additional thematic sessions that will cover particular topics more specialized.

We sincerely thank all the members of the Program Committee who helped us to set up a great scientific program by reviewing all submissions in time. This task would have been impossible without them! We also are grateful to the six invited speakers that have accepted to contribute to the success of the JOBIM edition in Nantes.

We are indebted to the organizing institutions, the SFBI, the GDR BIM, and the IFB. We are also grateful to all our partners and sponsors for their financial support.

Finally, we could also not forget to warmly thank Sophie Girault, Elodie Guidon, Aurore Morvan, and Jérémie Ségard as well as all the members of the organizing committee who worked collectively without counting sweat and tears to welcome the cream of bioinformaticians today in the best conditions.

Damien Eveillard and Audrey Bihouée
Organizing committee

Jérémie Bourdon and Richard Redon
Program committee

Program committee

Chloé-Agathe Azencott
Benjamin Bardiaux
Anais Baudot
Benoit Bely
Séverine Bérard
Camille Berthelot
Jérémie Bourdon
Laurence Calzone
Samuel Chaffron
Madalena Chaves
Hélène Chiapello
Sarah Cohen-Boulakia
Erwan Corre
Alexandre G. De Brevern
Damien Eveillard
Christine Froidevaux
Olivier Gandrillon
Christine Gaspin
Pierre-Antoine Gourraud
Carrier Gregory
Carito Guziolowski
Carl Herrmann
Claudine Landès
Solena Le Scouarnec
Claire Lemaitre
Cedric Lhoussaine
Pierre Lindenbaum
Anne Lopes
Morgan Magnin
Anthony Mathelier
Macha Nikolski
Eric Pelletier
Guy Perrière
Ovidiu Radulescu
Richard Redon
Eric Rivals
Frédéric Saubion
Sophie Schbath
Aurélien Sérandour
Anne Siegel
Christine Sinoquet
Stéphane Téletchéa
Dominique Tessier
Denis Thieffry
Morgane Thomas-Chollier
Hélène Touzet
Jacques van Helden

Local organizing committee

Audrey Bihouée (l'institut du thorax)
Damien Eveillard (LS2N)
Isabelle Alves (l'institut du thorax)
Agnes Basseville (CRCINA)
Jérémie Bourdon (LS2N)
Karine Cantele (LS2N)
Grégory Carrier (IFREMER)
Samuel Chaffron (LS2N)
Eric Charpentier (l'institut du thorax)
Benjamin Churcheward (LS2N)
Hélène Chiapello (SFBI)
Erwan Delage(LS2N)
Christian Dina (l'institut du thorax)
Solenne Dumont (l'institut du thorax)
Axelle Durand (CRTI)
Guillaume Fertin (LS2N)
Adrien Foucal (l'institut du thorax)
Victor Gaborit (CRCINA)
Alban Gaignard (l'institut du thorax)
Matthieu Giraud (CRTI)
Sophie Girault (LS2N)
Pierre-Antoine Gourraud (CRTI)
Elodie Guidon (LS2N)
Carito Guziolowski (LS2N)
Julie Haguait (LS2N)
Géraldine Jean (LS2N)
Matilde Karakachoff (CHU)
Abdelhalim Larhlimi (LS2N)
Solena Le Scouarnec (l'institut du thorax)
Sébastien Leuillet (Biofortis)
Sophie Limou (CRTI)
Pierre Lindenbaum (l'institut du thorax)
Dimitri Meistermann (CRTI)
Stéphane Minvielle (CRCINA)
Jérémie Poschmann (CRTI)
Richard Redon (l'institut du thorax)
Jennifer Rondineau (CRCINA)
Bruno Saint-Jean (IFREMER)
Aurélien Serandour (CRCINA)
Floriane Simonet (l'institut du thorax)
Dominique Tessier (INRA BIA)
Raluca Teusan (l'institut du thorax)
Camille Trottier (LS2N)
Nicolas Vince (CRTI)

Our Partners and Sponsors

Invited Speakers

Chloé-Agathe Azencott

Chloé-Agathe Azencott is an assistant professor of the Centre for Computational Biology (CBIO) of MINES ParisTech and Institut Curie (Paris, France). She earned her PhD in computer science at University of California, Irvine (USA) in 2010, working at the Institute for Genomics and Bioinformatics. She then spent 3 years as a postdoctoral researcher in the Machine Learning and Computational Biology group of the Max Planck Institutes in Tübingen (Germany) before joining CBIO. Her research revolves around the development and application of machine learning methods for biomedical research, with particular interest for feature selection and the integration of structured information. She currently holds funding from ANR (Jeune Chercheur Jeune Chercheuse project SCAPHE) and ERC (Inovative Training Network MLFPM). She is also the co-founder of the Parisian branch of Women in Machine Learning and Data Science.

Alexander Bockmayr

Alexander Bockmayr is a full professor at the Department of Mathematics and Informatics of Freie Universität Berlin since 2004. He holds the chair for Mathematics in Life Sciences. From 1998 to 2004 he was a professor at Lorraine University, Nancy, and head of the MODBIO (Computational Models in Molecular Biology) project-team at LORIA and INRIA. His current research focuses on mathematical and computational methods for molecular systems biology. The main topics of interest are metabolic and regulatory networks. Special emphasis is on constraint-based methods, i.e., reasoning with constraints, where each constraint represents a piece of partial information on the structure or dynamics of the network under study. His mathematical background lies in discrete mathematics and optimisation, constraint and integer programming, and computational logic.

Alessandra Carbone

Alessandra Carbone is Professor of Computer Science at Sorbonne Université, she leads the Analytical Genomics team since 2003 and is the director of the Department of Computational and Quantitative Biology since 2009. Her group works on computational problems concerning the functioning and evolution of biological systems. Mathematical methods coming from statistics and combinatorics, as well as algorithmic tools are employed to study fundamental principles of the cellular functioning starting from genomic, metagenomic and structural data. The projects are all aimed at understanding the basic principles of evolution and co-evolution of molecular structures in the cell.

Olivier Delaneau

Olivier Delaneau is currently a SNSF professor in the department of Computational Biology of the University of Lausanne. His research focuses on two main topics. First, he aims at better characterizing the molecular mechanisms underlying the genetic variations that affect the expression of genes (aka eQTLs). To do so, his group analyses large population scale genomic datasets regrouping genetic variations, expression of genes (measured using RNA-seq) and activity of regulatory elements (measured using ChIP-seq). This part of his research largely relies on network modeling and causal inference. Second, he also aims at improving methods for the imputation of genotypes and haplotypes from large scale genomic data sets (aka as imputation). To do so, his group develops fast and accurate statistical methods, usually based on Hidden Markov Models, and applies them on the large genomic data sets regrouping hundreds of thousands of genomes. In 2008, he obtained a PhD in bioinformatics from the Conservatoire National des Arts et Métiers (CNAM) in Paris and went through two successive postdocs between 2009 and 2018; in the department of Statistics of the University of Oxford (UK) and in the department of Genetic Medicine and Development of the University of Geneva (Switzerland). Through the last ten years, he developed widely used genomics software packages such as SHAPEIT, FastQTL and QTLtools that were in large scale projects such as 1000 Genomes, Haplotype Reference Consortium, UK Biobank and GTEx.

Christophe Dessimoz

Christophe Dessimoz obtained his Master in Biology (2003) and PhD in Computer Science (2009) from ETH Zurich, Switzerland. After a postdoc at the European Bioinformatics Institute near Cambridge (UK), he joined University College London as lecturer (2013), then Reader (2015). In 2015, he joined the University of Lausanne as SNSF professor, retaining an appointment at UCL, where part of his lab remains active. Since 2016, Christophe is also a group leader at the Swiss Institute of Bioinformatics. At the interface of biology and computer science, Christophe's lab seeks to better understand evolutionary and functional relationships between genes, genomes and species. His lab develops and maintains the OMA (Orthology Matrix) resource. He is cautious proponent of a "Big Data" approach to bioinformatics.

Juliette Martin

Juliette Martin received her PhD in 2005 from University Paris 7, working at the Mathématiques Informatique et Génome Unit at INRA, Jouy-en-Josas. After a first post-doc in Paris at INSERM/Paris 7 and a second post-doc at the Indian Institute of Science in Bangalore, she joined the CNRS in Lyon as a full-time researcher in 2008. Her research focuses on the structural bioformatics of protein-protein interactions: prediction of interactions and interaction sites via information gained from structures.

Guy Cochrane

Dr Guy Cochrane leads the European Nucleotide Archive (ENA), a platform for the management, sharing, integration and dissemination of sequence data. ENA includes, on the technical side, core databasing infrastructure for the rapid archiving of petabytes of sequence data, submission/validation services used by several 1000s of data providers, and sophisticated data discovery and retrieval tools used by many times this number. On the content side, ENA offers extensive public domain data from over 2 million species. Providing the European node of the celebrated long-standing International Nucleotide Sequence Database Collaboration, Cochrane is an authority on large-scale international sequence data sharing across application areas and taxonomies. Within his current portfolio of projects, for example, his team leads on data coordination across marine, pathogen and livestock data coordination. Cochrane has driven numerous developments within sequencing informatics, notably data standards; global next generation sequence data infrastructure; CRAM sequence data compression software; the data hub and portal system; and most recently tools and services for data brokering.

Abstracts

Contents

Talks

UniFire: the UNIprot Functional annotation Inference Rule Engine

Alexandre RENAUX[1], Hermann Zellner[2], Maria MARTIN[2] and Rabie SAIDI[2]

[1] Interuniversity Institute of Bioinformatics in Brussels, Université libre de Bruxelles-Vrije Universiteit Brussel, 1050 Brussels, Belgium

[2] European Molecular Biology Laboratory, European Bioinformatics Institute (EMBL-EBI), Cambridge, United Kingdom,

Corresponding Authors: rsaidi@ebi.ac.uk / alexandre.renaux@ulb.ac.be

Abstract *UniFIRE (The UniProt Functional annotation Inference Rule Engine) is an engine to execute rules in the UniProt Rule Markup Language (URML) format. It can be used to execute the UniProt annotation rules (UniRule and SAAS). This project is a work in progress and is available at http://ftp/pub/contrib/UniProtKB/UniFIRE. We would like to work with the scientific community on this development and encourage users to register their interest in the links provided on our blog https://insideuniprot.blogspot.com/2018/03.*

Keywords Functional annotation, Rule Engine, Inference, UniProt, URML

With the increasing number of sequence data generated by high-throughput sequencing methods, biological researchers need reliable automatic systems to provide the functional annotation of predicted proteins. The Universal Protein Knowledgebase (UniProtKB) is using two automatic annotation systems, UniRule and the Statistical Automatic Annotation System (SAAS), to automatically annotate UniProtKB/TrEMBL in an efficient and scalable manner with a high degree of accuracy. These systems use protein signatures and taxonomy classifications to infer the biochemical features and biological functions of proteins. This knowledge is expressed in the form of rules: a set of IF-THEN statements coming from expert curation (UniRule [1]) or generated by machine learning (SAAS [1] and ARBA [2]).

The predicted annotations and their corresponding rules are publicly available yet the expert knowledge from these rules cannot be integrated and executed to annotate newly predicted protein sequences. The automatic annotation community could also benefit from those annotation systems, as some protein sequences may not be yet available in public databases or could be present in an highly redundant proteome absent from UniProtKB, thus not annotated by the UniProt annotation systems but whose sequences are present in the UniParc sequence archive dataset.

For these purposes, we have developed UniFIRE (the UniProt Functional Inference Rule Engine): an open-source Java-based framework and tool to apply the UniProt rules on given protein sequences. We also propose a well-defined rule format based on XML: URML (the UniProt Rule Markup Language), along with its corresponding data model, to facilitate the exchange and authoring of rules and to improve the interoperability and reusability of the UniProt knowledge on proteins. The tool we have developed is able to read the user's input data, match them to the UniProt rules and infer protein annotations. It embeds Drools, an open-source technology using an optimised version of the Rete algorithm to match facts and rules in a scalable way [3].

By using UniFIRE, the UniProt annotation systems have both been successfully leveraged in MicroScope, a prokaryotic annotation platform and we are reaching out for more successful collaborations in the future. We expect this rule format and engine will facilitate knowledge exchange and collaborations within the automatic annotation community.

References

[1] The UniProt Consortium, UniProt: a worldwide hub of protein knowledge, *Nucleic Acids Research*, Volume 47, Issue D1, 08 January 2019, Pages D506–D515, https://doi.org/10.1093/nar/gky1049.

[2] Rabie Saidi, Imene Boudellioua, Victor Solovyev and Maria Martin. Rule Mining Techniques to Predict Prokaryotic Metabolic Pathways. Methods Mol Biol Volume 1613 (2017) p.311-331.

[3] Mark Proctor (2012) Drools: A Rule Engine for Complex Event Processing. In: Schürr A., Varró D., Varró G. (eds) Applications of Graph Transformations with Industrial Relevance. AGTIVE 2011. Lecture Notes in Computer Science, vol 7233. Springer, Berlin, Heidelberg.

[4] David Vallenet, Alexandra Calteau, Stéphane Cruveiller, Mathieu Gachet, Aurélie Lajus, Adrien Josso, Jonathan Mercier, Alexandre Renaux, Johan Rollin, Zoe Rouy, David Roche, Claude Scarpelli, Claudine Médigue, MicroScope in 2017: an expanding and evolving integrated resource for community expertise of microbial genomes, Nucleic Acids Research, Volume 45, Issue D1, January 2017, Pages D517–D528, https://doi.org/10.1093/nar/gkw1101

"ProteoRE, a Galaxy-based platform for the annotation and the interpretation of proteomics data in biomedical research".

Florence COMBES[1], David CHRISTIANY[1,2], Virginie BRUN[1], Christophe CARON[2], Valentin LOUX[2], Yves VANDENBROUCK[1]

[1] Univ Grenoble Alpes, CEA, INSERM, BGE U1038, F-38000, Grenoble, France
2 MaIAGE, INRA, Université Paris-Saclay, Domaine de Vilvert, 78350, Jouy-en-Josas, France

Corresponding Author: yves.vandenbrouck@cea.fr; proteore@contact.fr

Summary

With the increased simplicity associated with producing MS-based proteomics data, the bottleneck has now shifted to the functional analysis and exploration of large lists of expressed proteins to extract meaningful biological knowledge. Bioinformatics resources are often spread and disseminated under different forms (program/libraries/software/web tools and databases) and their access is rather limited for researchers without programming experience or no in-house bioinformatics support. As a consequence, interpretation of their data by experts remains a tedious and time-consuming process, and potentially error-prone (e.g., due to manual handling or input error, use of outdated resources). The ProteoRE (Proteomics Research Environment) aims at fulfilling this need by centrally providing an online research service to assist biologists/clinicians in the interpretation of their proteomics data in a unified framework. Built upon the Galaxy environment, this web-based platform for computational biomedical research, allows researchers to apply a large range of dedicated bioinformatics tools and data analysis workflows on their data, share their analyses with others, and enable tiers to repeat the same analysis while keeping tracks of the overall process. Currently, ProteoRE implements 18 tools organized into four subsections for: i) data manipulation; ii) human and mouse species annotation; iii) functional analysis; and iv) pathway analysis along with graphical representations. Furthermore, we also developed a specialized tool that allow for the management (i.e. download and creation of data stored/indexed within the platform) of annotation from external resources upon which some ProteoRE's tools rely on (e.g. Uniprot, Human Protein Atlas, Biogrid, etc.). The update of these external resources on a regular basis allows to overcome the issue of outdated annotations while keeping alive previous releases to ensure the reproducibility in case of re-analysis. The ProteoRE platform is designed in collaboration with biomedical researchers and has recently been implemented for the functional analysis of a human MS/MS proteomics sample [1] and the selection of candidate proteomics biomarkers of human disease [2]. Since its opening in May 2018, around 2000 working sessions were performed by ~450 different users with a continuous progression (source: google analytics). Our platform also provides online support, tutorials and training material (soon shared via the Galaxy Training Network) and is in free access: http://www.proteore.org. In accordance with Galaxy's best practices, ProteoRE's tools are deposited in the Tool shed (https://toolshed.g2.bx.psu.edu/view/proteore) and we are open to any contribution or wishes that would enhance and/or broaden its analytical range. A brief introduction followed by a demo illustrating how ProteoRE can contribute to the field of biomedical research will be done.

Acknowledgements

This work was partly supported by grants from the "Investissement d'Avenir Infrastructures Nationales en Biologie et Santé" program (French Bioinformatics Infrastructure grant ANR-11-INBS-0013). We would like to thank the following for their contributions to the design and beta-testing of these tools: C. Bruley, B. Gilquin, M. Lacombe, L. Perus, M. Tardif. This work is dedicated to the memory of Christophe Caron.

References

1. Lacombe M et al. Proteomic characterization of human exhaled breath condensate. J Breath Res. 12(2):021001, 2018.
2. Nguyen L. et al., Designing an In Silico Strategy to Select Tissue-Leakage Biomarkers Using the Galaxy Framework. Methods Mol Biol. 1959:275-289, 2019.

Redesign of iPPI-DB, a database for modulators of Protein-Protein Interactions

Rachel TORCHET[1,†], Alexandra MOINE-FRANEL[1,2,3,†], Hélène BORGES[1,2,3], Bryan BRANCOTTE[1], Olivia DOPPELT-AZEROUAL[1], Fabien MAREUIL[1], Hervé MÉNAGER[1], and Olivier SPERANDIO[1,2,3*]

[1] C3BI, Institut Pasteur 28 rue du Dr Roux, 7015 Paris
[2] Structural Bioinformatics Unit, Institut Pasteur 28 rue du Dr Roux, 7015 Paris
[3] CNRS UMR 3528, Institut Pasteur 28 rue du Dr Roux, 7015 Paris

† These authors have equally contributed
Corresponding Author: olivier.sperandio@pasteur.fr

iPPI-DB, for inhibitors of Protein-Protein Interaction DataBase, is a web application first released in 2012, which stores physicochemical and pharmacological data about PPI modulators and their targets. Users can query the database using either pharmacological criteria or chemical similarity with a user-defined query compound. The database is manually curated from the scientific literature and contains more than a thousand non-peptide inhibitors (iPPI) across 18 families of Protein-Protein Interactions. In the initial version, The chemical structures, as well as the physicochemical and the pharmacological profiles of these compounds and their targets, were extracted from the literature, computed and retrieved using numerous manual steps. This rather tedious procedure was seriously hindering the updates of the database.

For this project, we applied a combination of Agile methods and a User-Centered Design (UCD) approach to completely redesign the iPPI-DB web application. The main goal of this redesign is to focus on the needs of the user, ensuring that the end product fits the purpose, increasing the number of entries in the database and easing the query process. We adopted an iterative approach, interleaving successive series of design, tests and implementation steps, involving users in each iteration. This process, although it required an important involvement from the users during the project, has been extremely beneficial, as it allowed us to build a constructive dialog between scientists and the development team, and quickly validate or ask for corrections in the software.

The resulting web application provides a rich, robust, and innovative software environment to facilitate the growth and the maintenance of the database, and to query it using a highly intuitive and extremely powerful user interface.

Keywords Protein-Protein interaction, Database, Web interface, UX Design

Introduction

Pharmaceutical innovation is still impaired by the paucity of clinically testable targets and by the fact that only a few are successfully exploited in each therapeutic area [1]. This stands in sharp contrast with the number and diversity of roles of Protein-Protein Interactions. Indeed, with about 130,000 binary PPIs and possibly more just in humans [2], the development of drugs targeting these systems, represents a significant step toward expanding the druggable genome [3] and a possible leverage on the pharmacological modulation of disease-associated cellular pathways.

Historically, the design of small molecular drugs targeting PPIs has been extremely challenging, such that it seems there is a pharmacological cost to pay when choosing such target: selecting the right PPI and the right drug chemotype to work with. Yet, a growing number of successful examples is demonstrating on a daily basis that such an endeavor is accessible when deploying the necessary means: solid knowledge of the biological pathways around the chosen target and extensive medicinal chemistry to identify and optimize new chemical probes. The recent approval of Venetoclax [4], as a small molecular drug targeting some of the anti-apoptotic members of the Bcl-2 family in specific types of Lymphoma, is the perfect example of this virtuous combination that can be learned from.

In the light of this context, there is great value in storing in organized databases the knowledge coming from such studies. It is the purpose of several consortia such as ChemBL [5] or Pubchem [6]. In the field of PPIs, two databases have paved the way either by automatically deriving chemical data

from ChemBL like TIMBAL [7], or by focussing on co-crystallized compounds like the 2P2I database [8]. But none of them has a thorough modeling of the data, nor an elaborated web application to query them.

When we developed iPPI-DB, we decided to use a complementary approach vis a vis the existing iPPI databases. First, we chose to manually store a substantial number of metadata about the PPI targets, the chemical compounds and their activities, as well as the experimental assays that produced those activities. Second, we designed an intuitive web application to allow users to efficiently access the desired data. We first reported iPPI-DB in 2013 [9] and significant improvements were subsequently made to add more PPI modulators and targets, as well as a chemical similarity query mode [10].

In those initial versions of the database, the addition of new data was fastidious.

The new version of iPPI-DB has been created to ease the addition of new entries through a convivial interface and to improve query capabilities for data retrieval. The resulting database is available through a powerful web application that will enable users to query and navigate the contents of the database in a multitude of ways, but also a contribution wizard that guides them through the process of suggesting new entries.

We here describe the organisation and approaches we adopted during the project, focusing on two points in particular: the project management, and the user centered design methodologies we used.

2. Project management and coordination

The size and ambition of the iPPI-DB project required the combined efforts of a research group and a software engineering team, mobilizing an important number of different expertises over the course of two years. We describe here the main guidelines that we adopted to facilitate the development of this new version, which can all be linked to the Agile methods [11], a set of practices that have been increasingly adopted in Bioinformatics software development [12]. This approach focuses on collaboration, communication, and interaction between the different stakeholders.

2.1. Iterative approach

Given the complexity of this project, which includes contributions from experts in Structural Bioinformatics, Software and Database Development, User Interface Design, we structured the project around an iterative approach, interleaving successive series of design, software development, and user tests. Such iterations initially focused on specific topics, such as the analysis of the existing version of iPPI-DB, the redesign of the database, or the design or the web interfaces. This approach, although it required a significant involvement from the researchers, has been extremely beneficial, as it allowed us to build a constructive dialog between scientists and the development team, and quickly validate or correct the software when needed.

2.2. Supporting infrastructure

To support the development of this project, we heavily relied on the infrastructure provided by the IT department. It includes (1) a gitlab server that provides version control and sharing for the source code of the application and other capabilities [13] such as issue tracking and release management, (2) a virtual machine that hosts the system, (3) and a GitLab CI/CD server to automatically run tests and deploy the latest version. This infrastructure enables:

- Collaborative software development, enabling all partners to access the source code and contribute through source modifications or issue reports,
- Quality monitoring through continuous testing
- Automated deployment of new versions.

This infrastructure enabled us to adopt "DevOps" practices[1] to build and share the source code and accelerate the deployment of corrections and new features.

3. User-Centered Design Approach and Technical architecture

We adopted a User-Centered Design (UCD) approach where the needs of users are taken into account all along the project. During the early stages of this process, we focus on understanding user

[1] https://en.wikipedia.org/wiki/DevOps

behaviors, needs, and goals. This results in the identification of three kinds of users: (1) The common users who for instance search for compounds based on chemical similarity or PPI target; (2) The external contributors who suggest new entries based on data from in publications; and (3) The core curators who both enter new data and validate external contributions. Along with these different types of users we defined different goals, expectations, and needs. We used specific UX methodologies to answer questions and to design the different interfaces.

3.1. Query interface revisited

The query interface allows selecting and visualizing the different compounds available, based on biological, chemical, and pharmaceutical criteria. Based on the needs expressed by the users, providing a convivial and efficient interface was mandatory to extract the best of the available data.

We invited the users to *Six Up and One Up* workshops [14], to design mockups and prototypes for the different pages. During such workshops, each participant receives six templates of an empty screen and has to draw six different versions of the user interface. These different prototypes are then presented and compared. This allows identifying the redundant functionalities and needs, the eventual pain points and to bring up new design questions. As a result, all prototypes are summarized in a final accurate version. This method allows us to create a consensus prototype within two meetings. Although this approach has been previously applied in some bioinformatics projects [15], it remains largely unusual. The process is easy to set up with biologists and engineers and is highly effective. Using this methodology helps to generate many ideas over a short time, and gives to all participants the opportunity to contribute. Additionally, since these query capabilities already existed in the previous version of iPPI-DB, providing many different points of view enabled us to avoid retaining the same user interface with which many participants (but not all) were already familiar.

The revisited query interface now lets users select, filter and visualize the different compounds available, based on biological, chemical, and pharmaceutical criteria. It also allows to refine and combine multiple filters to build complex queries, share them easily as URLs with collaborators, and download corresponding data. Query results can be displayed with different layouts (thumbnails, list, or table), and all of them can be sorted according to different parameters.

3.2. New contribution interface

iPPI-DB was developed as a manually curated database from the scientific literature that contains the structure, some physicochemical characteristics, the pharmacological data and the profile of the PPI targets of several hundred modulators of protein-protein interactions. The main limitation of this system is the addition of new entries: the full process relies on several disconnected scripts, different languages and processes, namely in R, Java, Perl, python, starting from simple data sheets compiling the data. This makes the update process complex and error-prone.

We designed a new contribution interface to ease this task, in close collaboration between the developers and the users. To that end, we ran prototyping meetings with users, in order to create a convivial interface which uses a step-by-step approach to lower the workload for the experts. During these focus groups, we discussed openly between all the participants about the functionalities of the interface to provide simple, scaled down versions of it. The interfaces were designed as wireframes, and later refined as interactive prototypes, which users tested to validate the usability of the interface. We iterated our design through different rounds of tests.

The resulting *contribution interface* is wizard-based, i.e, it is a succession of screens that guide users to enter the data needed to populate the database. Users provide the architecture of the PPI complex(es), the chemical compounds tested for modulation, and the various assays in which those compounds were tested. The interface requests minimal participation from users to reduce the risks for errors and facilitate contributions: whenever contributors provide some information, the server automatically retrieves additional details from other reference databases, such as Pubmed, Uniprot, or the PDB.

Conclusion

The upcoming iPPI-DB web application provides a rich, robust, and innovative software environment to facilitate the growth and the curation of the database, as well as to query it using a highly interactive and powerful user interface.

Acknowledgements

We wish to thank the IT department of the Institut Pasteur, especially Eric Deveaud, Thomas Menard, Jean-Baptiste Denis, Emmanuel Guichard and Youssef Ghorbal, as well as Quang Tru Huynh from the Structural Bioinformatics unit for helping us set up the infrastructure for this project. We also thank Chemaxon (http://www.chemaxon.com) for providing us with JChem (JChem 19.8.0, 2019), and MarvinsJS (Marvin 19.13.0, 2019) licenses.

References

1. Teague SJ. Learning lessons from drugs that have recently entered the market. Drug Discov Today. 2011;16: 398–411. doi:10.1016/j.drudis.2011.03.003

2. Venkatesan K, Rual J-F, Vazquez A, Stelzl U, Lemmens I, Hirozane-Kishikawa T, et al. An empirical framework for binary interactome mapping. Nat Methods. 2009;6: 83–90. doi:10.1038/nmeth.1280

3. Laraia L, McKenzie G, Spring DR, Venkitaraman AR, Huggins DJ. Overcoming Chemical, Biological, and Computational Challenges in the Development of Inhibitors Targeting Protein-Protein Interactions. Chem Biol. 2015;22: 689–703. doi:10.1016/j.chembiol.2015.04.019

4. Leverson JD, Phillips DC, Mitten MJ, Boghaert ER, Diaz D, Tahir SK, et al. Exploiting selective BCL-2 family inhibitors to dissect cell survival dependencies and define improved strategies for cancer therapy. Sci Transl Med. 2015;7: 279ra40. doi:10.1126/scitranslmed.aaa4642

5. Gaulton A, Hersey A, Nowotka M, Bento AP, Chambers J, Mendez D, et al. The ChEMBL database in 2017. Nucleic Acids Res. 2017;45: D945–D954. doi:10.1093/nar/gkw1074

6. Wang Y, Bryant SH, Cheng T, Wang J, Gindulyte A, Shoemaker BA, et al. PubChem BioAssay: 2017 update. Nucleic Acids Res. 2017;45: D955–D963. doi:10.1093/nar/gkw1118

7. Higueruelo AP, Jubb H, Blundell TL. TIMBAL v2: update of a database holding small molecules modulating protein-protein interactions. Database . 2013;2013: bat039. doi:10.1093/database/bat039

8. Basse M-J, Betzi S, Morelli X, Roche P. 2P2Idb v2: update of a structural database dedicated to orthosteric modulation of protein-protein interactions. Database . 2016;2016. doi:10.1093/database/baw007

9. Labbé CM, Laconde G, Kuenemann MA, Villoutreix BO, Sperandio O. iPPI-DB: a manually curated and interactive database of small non-peptide inhibitors of protein–protein interactions. Drug Discov Today. 2013;18: 958–968. doi:10.1016/j.drudis.2013.05.003

10. Labbé CM, Kuenemann MA, Zarzycka B, Vriend G, Nicolaes GAF, Lagorce D, et al. iPPI-DB: an online database of modulators of protein–protein interactions. Nucleic Acids Res. Narnia; 2016;44: D542–D547. doi:10.1093/nar/gkv982

11. Kane DW, Hohman MM, Cerami EG, McCormick MW, Kuhlmman KF, Byrd JA. Agile methods in biomedical software development: a multi-site experience report. BMC Bioinformatics. 2006;7: 273. doi:10.1186/1471-2105-7-273

12. Prlić A, Procter JB. Ten simple rules for the open development of scientific software. PLoS Comput Biol. 2012;8: e1002802. doi:10.1371/journal.pcbi.1002802

13. Perez-Riverol Y, Gatto L, Wang R, Sachsenberg T, Uszkoreit J, Leprevost F da V, et al. Ten Simple Rules for Taking Advantage of Git and GitHub. PLoS Comput Biol. 2016;12: e1004947. doi:10.1371/journal.pcbi.1004947

14. Bowles C, Box J. Undercover User Experience Design [Internet]. Pearson Education; 2010. Available:

https://market.android.com/details?id=book-9fwTj1japZcC

15. Karamanis N, Pignatelli M, Carvalho-Silva D, Rowland F, Cham JA, Dunham I. Designing an intuitive web application for drug discovery scientists. Drug Discov Today. 2018;23: 1169–1174. doi:10.1016/j.drudis.2018.01.032

A review of different ways to insert known RNA modules into RNA secondary structures

Louis Becquey, Eric Angel and Fariza Tahi
IBISC, Univ Evry, Universite Paris-Saclay, 91025, Evry, France

Corresponding author: louis.becquey@univ-evry.fr

Abstract *A common approach to RNA folding pipelines is to start by predicting secondary structures from sequence, and then, tertiary spatial folds from the secondary structure. In this review, we are interested in a backward approach: we explore what information from known solved RNA 3D structures can be used to improve secondary structure prediction. We propose a Pareto-based method for predicting secondary structures by minimizing a bi-objective half energy-based, half knowledge-based potential. The tool outputs the secondary structures from the Pareto set. We use it to compare several approaches to insert RNA modules into the secondary structures and benchmark them against the RNAstrand secondary structure database. We compare two different module data sources, Rna3Dmotif and The RNA Motif Atlas and different ways to score the module insertions taking into account module size, module complexity, or module probability according to models like JAR3D and the recent BayesPairing method.*

Keywords RNA modules, secondary structure, integer programming, benchmark, multi-criteria optimization

1 Introduction

Ribonucleic acid (RNA) is a macromolecule which is often single-stranded. Therefore, the strand has the ability to fold in space in more complex ways than DNA, that we mostly know to form double-stranded stems. A stem is a succession of basepairs called Watson-Crick basepairs, or "canonical", stacked on top of each other. As this can still happen with RNA, we also observe several other ways for a nucleotide to interact with another. For example, Leontis and Westhof [1] proposed a classification of 12 non-canonical basepairs. Some of the nucleotides can also interact with the 2'OH of a ribose, or with a phosphate, or even not interact at all and bulge out the RNA structure.

Modelling an RNA as a graph. For modelling purposes, researchers working on computational problems involving RNA represent them with graphs. A recent article by Schlick [2] details the different graph models of RNAs and their respective advantages. We are particularly interested in the secondary structure graph of the RNA, i.e. a graph where the nucleotides are nodes, and backbone bonds and canonical basepairs are edges. In this kind of graph, the non-canonical interactions do not appear. As the problem of predicting the 3D structure of an RNA from sequence as been too computationally expensive for years, and is still difficult, a common first step has been to predict this secondary structure (2D) graph, by computing what regions will form stems and what regions will remain unpaired, forming so-called loops. Note that non-canonical interactions are not considered in the secondary structure graph. In many cases, the solution to the 2D folding problem is not unique, and RNAs have the ability to switch between several stable conformations. Most approaches are based on the computation of the canonical pairing probabilities, i.e. the probability for each nucleotide to form a canonical base-pair with every other nucleotide, or to remain unpaired. In 1990, McCaskill proposed a dynamic programming scheme to compute those probabilities [3]. The most used implementations are some lower complexity variants of it such as `RNAFold` in the ViennaRNA package [4], `Fold` or `ProbKnot` from the RNAstructure package [5,6], and a variant taking pseudoknots into account from the NUPACK package [7]. Once those probabilities are computed, several models exist to rebuild one or several best structure(s): We can choose the Minimum Free Energy (MFE) structure, the one that maximizes expected accuracy (MEA), or the centroïd of the ensemble. We can also cite Biokop [8], a recent tool that uses both MFE and MEA criterions in a biobjective framework, and returns optimal and suboptimal structures.

Modelling loops with more detailed graphs. Then, the modeller needs to move from the planar 2D graph to 3D. Stems are relatively easy to tackle because of the isostericity of the Watson-Crick basepairs, their structure has been widely observed and features low variability. On the other hand, to accurately model loops in 3D, one needs to take the non-canonical interactions into account. Several works have gathered 3D crystal structures involving RNA chains, then extracted the loops from those chains and annotated the nucleotide contacts using MC-Annotate [9], FR3D [10] or DSSR [11] to model the loops with more detailed graphs with edges describing non-canonical contacts. The graphs can then be clustered with respect to a similarity or isomorphism measure, and the sequence variations over the nucleotides of the loop can be modeled. Those models are called RNA *modules*, i.e. an ordered collection of non-canonical basepairs or stacking interactions, leading to a conserved 3D shape in different RNA molecules.

We can cite the work from Djelloul and Denise [12] with Rna3Dmotif, a pipeline that extracts terminal hairpin loops, internal loops, and multiple loops from structures annotated by FR3D, and can cluster them using a graph similarity metric. Another one is The RNA Motif Atlas [13], which does not support multiple loops, but clusters the loops using all sequence, nucleotide contacts and shape information, which leads to loop module models with tolerance in sequence and length variations. A more recent one is CaRNAval [14], an approach that enables to model a wide variety of structural features such as multipairs, multi-stranded loops, and pseudoknots. To be exhaustive, we also can cite RNA Bricks 2 [15], which has the particularity to also study contacts with protein chains, and RNA Motif Scan [16] that can search for certain modules in structures, but does not list modules on the form of a database. These methods provide module models combining different types of information: (i) a particular base-pairing pattern of canonical and wobble pairing, which is 2D information, and can be limited to the canonical base-pairs that enclose a loop; (ii) a particular organisation of non-canonical contacts in space, which is 3D information; (iii) a sequence or consensus sequence that we know to adopt a particular base-pairing organisation, which could be nucleotide probabilities observed in the training dataset of RNA structures, or a more elaborated probabilistic model to predict if a given sequence will fold according to the module. For example, JAR3D [17] can score the modules from the RNA Motif Atlas against a query sequence. The recent BayesPairing [18], expanding the method proposed by Cruz and Westhof with RMDetect [19], can be used to do the same on modules from any database by building Bayesian networks from any graph of ordered non-canonical interactions.

Motivation of this work. Here, we are not interested in using the module models to rebuild 3D structures, but instead we want to see if that information could be used to predict the position of loops in sequences, in other words, if this data could help predicting the secondary structure graph.

A first attempt to tackle such task, called RNA-MoIP [20], corrects an input secondary structure to insert modules from Rna3Dmotif into it. The authors have shown that RNA-MoIP produces 2D structures which are better inputs to give to MC-Sym [21], resulting in better prediction of 3D structures. Unfortunately, we were not able to reproduce the published results about 2D structure accuracy, and our tests over 590 structures from the RNA Strand database [22] showed inferior performance compared to RNAsubopt [4], as shown in figure 1. It seems that RNA MoIP damages the structures predicted by RNAsubopt most of the time, leading to a slightly weaker performance.

One hypothesis about RNA-MoIP's lack of performance is that it cannot distinguish important base-pairs from less important ones, and might break some of the ones stabilizing a whole stem while inserting a module, resulting in lower probable structures as output. To test this hypothesis, we design a method which builds a 2D structure by simultaneously placing base-pairs and modules in a single step, taking into account two objectives: the expected accuracy of the structure in the equilibrium ensemble fold, and a custom function that reflects the number and quality of inserted modules (several models are studied).

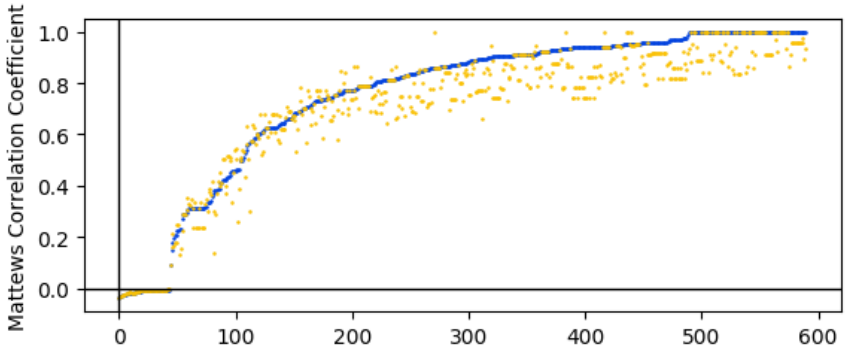


Fig. 1. Best Mattews Correlation Coefficient found between the true structure and the structures predicted by RNAsubopt (blue) and RNA MoIP (yellow) for 590 RNAs of length 10 to 100 from RNAstrand, sorted by RNAsubopt performance.

2 Methods

The main procedure we are using is the following:

— **Pattern-matching step:** Find all possible occurrences of known RNA modules in the query sequence, by finding subsequences of the query that score well with the probabilistic models of the modules (several models are compared).

— **Constraints building step:** Define constraints on the secondary structure imposed by modules if they would be included (in this case, some of the canonical base-pairs are forbidden).

— **Optimization step:** Find a secondary structure that satisfies as much as possible both the expected accuracy of the structure and a criterion taking into account module inclusions, by solving a bi-objective integer linear programming problem, using the previous constraints defined in the previous step.

The linear integer programming framework used to define the constraints and solve the resulting optimization problem is similar to previous works like IPknot, Biokop or RNA-MoIP.

2.1 Data sources

We compared two databases of modules: (1) Modules extracted from solved 3D RNA structures, in the DESC file format of Rna3dDmotif. We used the dataset provided by RNA-MoIP [12,20], (2) Modules from the RNA 3D Motif Atlas v3.2, as provided on the BGSU website [13].

All the RNA secondary structures used for the benchmarks are extracted from the RNA Strand database [22]. We selected the structures with size varying between 10 and 100 nucleotides. Sequences containing consensus letters, for example R for a purine (A or G), or modified nucleotides were discarded. The final dataset contains 590 secondary structures, with 97 containing pseudoknots.

2.2 Pattern matching step

Several methods have been proposed to tackle the issue of finding if a sequence (or a part of it) is likely to fold following a given module. This section presents the ones we benchmarked.

The SCFG/MRF method For each motif group in each release of the Motif Atlas, the BGSU RNA research group proposed to construct a probabilistic model for sequence variability, based on a hybrid Stochastic Context-Free Grammar/Markov Random Field (SCFG/MRF) method. Their implementation, a software called JAR3D [17], takes user-provided loop subsequences in input and outputs a score for every motif of the Atlas on every provided loop. This method has the advantage to allow variations in sequence length compared to the original module model. But it can only be used for hairpin and internal loops, and requires the computation of the pairing probabilities to first locate *where* the most probable loops are, to give them as input to JAR3D. This drawback is important, because we score modules on sequence portions that we already know unlikely to form stems. Therefore, the information brought by the insertion of a module is low. This method has only be tested on modules from the RNA 3D Motif Atlas.

Sequence probability distribution and Bayesian Networks When we have data for several instances of a module, we can estimate a probabilistic distribution of the nucleotides over the module nodes. A first intuitive approach is to use the base frequencies. But as paired nucleotides are not independant at all, it is more rigorous to model those dependancies. An approach proposed in [19] is to transform the module's graph into a bayesian network, which models the dependencies between nucleotide probabilities at every node of the graph. The original article proposed four hand-made bayesian networks for four well-known RNA modules, but a very recent BayesPairing software [18] automates the process for every module. A large number of sequences are sampled using the bayesian network, and they are pattern-matched against the query to find occurrences. An additional step compares the free energy of the structure with and without the constraint of each matched module, and selects only the candidate sites that do not deteriorate too much the energy. But this step requires two computations of the partition function, and leads to the same drawback than JAR3D: we try to insert modules that were pre-selected to be appropriate. Therefore we chose to ignore this last step and let our optimizer select the pertinent modules in the candidates. BayesPairing can be used for several data sources.

Direct pattern matching The simplest approach when no statistical model is available is to use a regular expression and direct pattern matching against the input sequence. This is the approach used by RNA-MoIP. We used it with the Rna3Dmotif data as presented in RNA-MoIP's article [20], dealing the same way with special cases (very short components, wildcards).

2.3 Constraints definition step and integer programming model

Here we propose different objective functions to maximize, whose performances are compared in section 3.

Notations Let x be a module which could be inserted at some defined position in the sequence. Let $\|x\|$ bet the number of components of this module, and $k_{x,i}$ the nucleotide count of the ith component of x. When a scoring model is used (JAR3D or BayesPairing), we denote $p(x)$ the score value of x inserted at the defined position. Let p_{uv} be the probability for nucleotides u and v (with $v > u+3$) to form a canonical base-pair. We use Dirks & Pierce's dynamic programming scheme [7], which supports pseudoknots, to compute such probabilities. We denote y_v^u the binary decision variable indicating that these nucleotides do form a canonical base pair, and C_1^x the decision binary variable indicating whether the module x will be inserted or not. The resolution of the linear program outputs solutions by fixing definitive values for the different y_v^u and C_1^x.

Objective functions The more modules that are included, the more information about set and unset base-pairs, and the more information we have about the tertiary folds of these loops in space. So maximizing the number of modules could be a valid criteria. But, a disadvantage of such a criteria is that it penalizes multiple loops - sometimes referred as k-way junctions - with large k, because the insertion of a multiple loop forbids at the same time the insertion of several internal loops or bulges (2-way junctions) in place. Then, we also want to maximize the number of components k in the module. This leads to our first criteria f_{1A}.
Then, we suppose that secondary structure contacts are local, and want to avoid very-long-range base-pairs. This is equivalent to say that we want the minimal loop size. Therefore, we can penalize a module insertion by the logarithm of the number of nucleotides involved in the looped zone (sum of the $k_{x,i}$) to avoid long unpaired zones. We introduce such a penalty in criteria f_{1B}.
We also define two more criteria, which use only the score returned by JAR3D or BayesPairing for f_{1C}, and all of the presented terms for f_{1D}.
Let X be the set of all our decision variables, then the different objective functions to maximize are:

$$f_{1A}(X) = \sum_x \sum_{i=1}^{\|x\|} k_{x,i}^2 \times C_1^x \qquad f_{1B}(X) = \sum_x \left[\frac{\|x\|}{\log_2(\sum_{i=1}^{\|x\|} k_{x,i})} \times C_1^x \right]$$

$$f_{1C}(X) = \sum_x p(x) \times C_1^x \qquad f_{1D}(X) = \sum_x \left[\frac{\|x\|}{\log_2(\sum_{i=1}^{\|x\|} k_{x,i})} \times p(x) \times C_1^x \right]$$

RNA-MoIP uses f_{1A} [20]. As first proposed by the IPknot authors [23], we use $f_2(X) = \sum y_v^u \times p_{uv} \times I[p_{uv} > \theta]$ as a second objective to maximize the expected accuracy of the secondary structure, using a parameter θ to ignore very unlikely base-pairs. This prevents the explosion of the number of variables and allows a fast resolution of the IP problem.

2.4 Optimization step

We use a simple dichotomic search algorithm (presented in Figure 2) to find the Pareto set of the bi-objective problem. It solves iteratively a mono-objective problem with a constraint on the second objective, requiring it to be in an interval $[\lambda_{min}, \lambda_{max}]$. Everytime a new non-dominated solution is found, λ_{min} is set just above the solution's objective 2 value, to search another one on top of it with a higher objective 2 value. Depending on our experiments, we used function f_{1A}, f_{1B} or f_{1C} as f_1. The expected accuracy criteria is f_2 as described above. The second pass of the dichotomy which searches *below* $f_2(s)$ is required to search for superposed solutions to Pareto optimal ones. This is important when the criteria used to rank inserted modules is not able to separate them very well; many solutions therefore get the same f_1 score.

The algorithm is implemented in C++ using the CPLEX solver concert technology [24].

Algorithm 1: FindParetoSet()

```
F:= ∅
L1:= maximize(f1, −∞, +∞, F)
L2:= maximize(f2, −∞, +∞, F)
R:= {L1}
// search on top of L1:
search_between(f2(L1) + ε, f2(L2))
search_between(−∞, f2(L1))
return R
```

Algorithm 2: search_between(λ_{min}, λ_{max})

```
s:= maximize(f1, λmin, λmax, F)
if s ≠ ∅ then
    F:= F ∪{s}
    if ∄x ∈ R such as x > s then
        R:= R ∪{s}
        while ∃x ∈ R such as s > x do
            R:= R\{x}
        end
        search_between(f2(s) + ε, λmax)
        if λmax − λmin > ε then
            search_between(λmin, f2(s))
        end
    end
end
```

Fig. 2. The dichotomic search algorithm to find the Pareto set. F is the ensemble of already-found structures which grows over time, and that we forbid the solver to find again. R is the set of pareto-optimal solutions. L1 and L2 are the best solutions to the mono-objective problems regarding $f1$ and $f2$. **maximize**(f, λ_{min}, λ_{max}, F) is a procedure that minimizes the function f (mono-objective IP problem) under the constraint that the other one has to be in interval $[\lambda_{min}, \lambda_{max}]$, and with the solutions in F forbidden. The inequality sign $a > b$ between two solutions denotes that solution a dominates solution b.

2.5 Additional compared methods

To study the usefulness of the data sources, objective functions, and module placement methods, we added state-of-the art tools to the comparison. The same RNAStrand sequences were submitted to RNA-MoIP for direct performance comparison. RNAsubopt (no pseudoknot support) and Biokop (bi-objectif integer programming framework with pseudoknot support) were added to the benchmark, both with default parameters.

3 Results

All the methods introduced return an ensemble of possible secondary structures for a given input sequence. We compute the Matthews correlation coefficient (MCC) between the real secondary structure and every proposition. Then, we keep the best MCC value found as a metric of the method's performance. The choice of MCC over accuracy or F1 score is justified by the very large difference

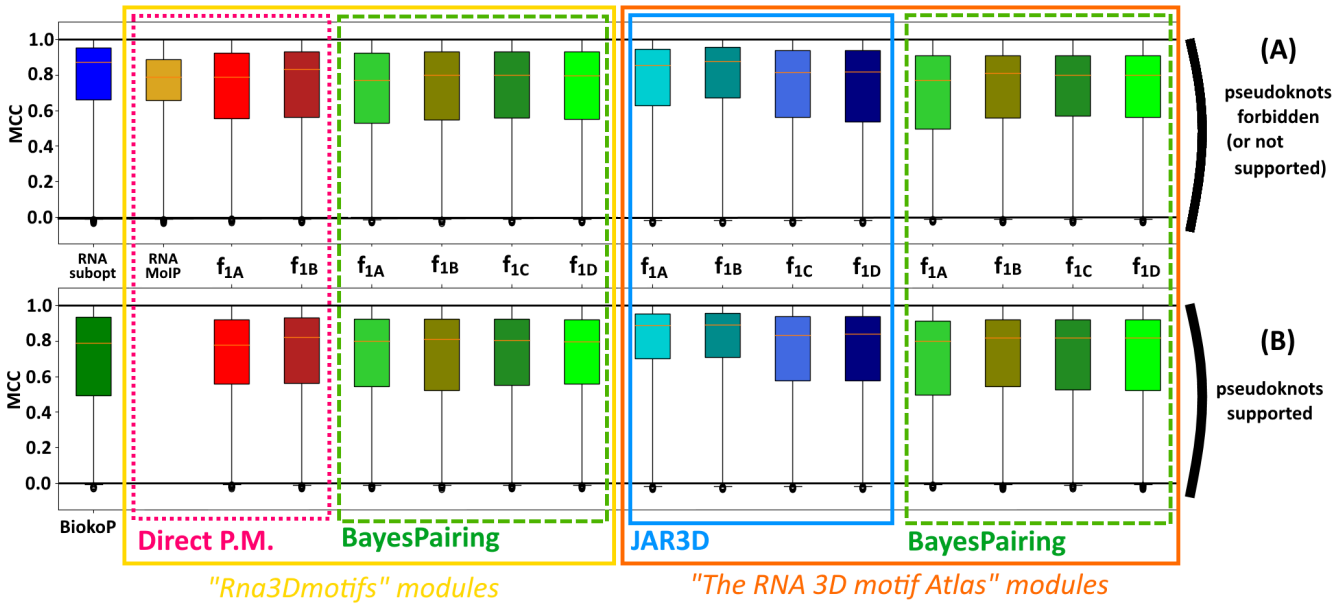


Fig. 3. Boxplots of the best MCC over the proposed solutions for each of the RNAs, for all method variants. The top line shows the methods that cannot find pseudoknots: RNAsubopt, RNA-MoIP, and the 14 variants of bi-objective methods with a constraint that explicitly forbids pseudoknots. The bottom line shows methods which allow their prediction: Biokop, and the 14 variants without the no-pseudoknot constraint. The left block gathers methods which use module data from Rna3dmotifs [12]. The right one gathers those which use modules from the RNA 3D Motif Atlas [13]. Boxplots surrounded by a dotted red frame use direct pattern matching to detect insertion sites, but do not score the sites. Those surrounded by a continuous blue frame score the sites with JAR3D [17] to score modules on loop sequences found by RNAsubopt. The remaining surrounded by a dashed green frame use the BayesPairing [18] score.

between the size of the classes: there exist much more negative base-pairs (pairs of nucleotides that do not interact) than positive ones in any secondary structure.

3.1 General benchmark results

Performance results under the form of best MCC are summarized in Figure 3. Majority of the RNAs were predicted with similar performance among the methods, including methods that do not use module information. Therefore, we can argue that including known modules is not a general way to improve secondary structure prediction; for every method, the performance gain obtained on some structures is counterbalanced by the loss on approximately the same number of RNAs.

Regarding the two module models, no data source, or module model, performs significantly better than the other one when looking only at the data source.

Regarding objective functions to include modules, the different criteria proposed seem to give comparable results regarding the average performance and the dispersion. However, an important difference between f_{1A}, f_{1B} on one side, and f_{1C}, f_{1D} on the other side, is about the number of solutions found in the Pareto set. As f_{1A}, f_{1B} do not use a score to rank potential module insertion sites, every modules of the same size can be equally inserted. When the RNA presents several loops, the combinatorial possibilities grow fast with the number of modules in the dataset. Therefore, the number of undominated solutions can reach several hundreds or thousands even for short sequences. For that reason, our computations for f_{1A} and f_{1B} with JAR3D never ended for 155 and 168 structures respectively. Such large Pareto sets are not informative for our application, because they consist in very redundant secondary structures with different module references, which are counted only for one solution at the end.

We also observe that using The RNA Motif Atlas with JAR3D has a significantly different behavior than the other methods: first, it returns a very small number of solutions (between 1 and 5 most of the time while other can often return from 10 to 50 solutions). Then, the best structure is almost everytime the one that has the higher number of modules, while it is not the case for the other methods. An

explanation is that JAR3D is selective of a few module insertion sites, sites that were first perfectly predicted to be loops by RNAsubopt (as discussed earlier in section 2.2). This confirms the use of module information is not relevant and the energy criteria brings almost all the information.

Without pseudoknots, comparison to RNAsubopt and RNA-MoIP The large increase in variance between boxplots of our methods compared to RNA-MoIP is the consequence of both the module information added and the bi-objective framework. RNA-MoIP adjusts loops predicted by RNAsubopt to include modules, which explains why it is always very close to RNAsubopt, while the bi-objective versions decides which modules and which base-pairs to keep in one run. It improves some predictions and deteriorates as many other ones.

Some loss of performance compared to RNAsubopt and RNA-MoIP can be explained by the fact the bi-objective methods are able to discard some of structures that would be proposed by RNAsubopt. Sometimes, this is an improvement, because it reduces the user's need to guess the right one. On the other hand, it also probably discards the best candidate sometimes, to insert a false-positive module.

With pseudoknots, comparison to Biokop Most of the RNAs are predicted with small knots when the method allows it. But we also notice that overall methods, this does not significantly increases nor reduces the performance. However, pseudoknot prediction quality is difficult to assess with a metric like MCC, because a pseudoknot could be involved in only two or three basepairs. Finding them or not does not alter much the MCC even if the structure is much more correct from a biological point of view. As the bottom line of Figure 3 shows, Biokop also predicts structures with the same performance magnitude than the bi-objective methods, and without module information.

4 Conclusion

In this review, a general bi-objective method was developed to benchmark different sources of RNA module models (the RNA 3D Motif Atlas and Rna3Dmotifs), different methods to place them in sequences (direct pattern matching, BayesPairing, and JAR3D), and different scoring functions. The biobjective method uses the expected accuracy of the structure, and the previous scoring functions to select relevant secondary structures.

The results show that objective functions which use a score on the module insertion site (produced by BayesPairing or JAR3D) do not lead to better accuracy than those which don't, but require much less computational resources to achieve the computation, as they avoid combinatorial explosion of the number of possible insertions of equally ranked modules onto the different loop sites of the RNA. The results show also that no data source prevails.

Some combinations overperformed RNA-MoIP, a previous attempt to predict better secondary structures using module information from Rna3Dmotifs and a linear combination of two objectives into a scoring function. But the general performance of these methods is below or equal to what RNAsubopt or Biokop can achieve without any module information. One of the best performing combination over the benchmarked methods is the use of Rna3dmotifs directly placed in the sequences using pattern-matching, and the presented bi-objective integer programming framework. This method could be interpreted as an upgraded RNA-MoIP with updated data (there is a 10-fold increase in the number of solved RNA crystal structures between 2008 and 2018) and a real bi-objective framework, which predicts the base pairs and the module insertions in a row. Another major interest of this method over RNAsubopt is the ability to predict pseudoknots, with faster computation times than Biokop.

Improvement perspectives now rely on the hope than newer databases like CaRNAval, containing more recent and more diverse module information, really bring some relevant information to assist the energy criteria.

References

[1] Neocles B Leontis and Eric Westhof. Geometric nomenclature and classification of RNA base pairs. *Rna*, 7(4):499–512, 2001.

[2] Tamar Schlick. Adventures with RNA graphs. *Methods*, 2018.

[3] John S McCaskill. The equilibrium partition function and base pair binding probabilities for RNA secondary structure. *Biopolymers: Original Research on Biomolecules*, 29(6-7):1105–1119, 1990.

[4] Ronny Lorenz, Stephan H. Bernhart, Christian Höner zu Siederdissen, Hakim Tafer, Christoph Flamm, Peter F. Stadler, and Ivo L. Hofacker. ViennaRNA Package 2.0. *Algorithms for Molecular Biology*, 6:26, November 2011.

[5] David H Mathews. Using an RNA secondary structure partition function to determine confidence in base pairs predicted by free energy minimization. *Rna*, 10(8):1178–1190, 2004.

[6] Stanislav Bellaousov and David H Mathews. Probknot: fast prediction of RNA secondary structure including pseudoknots. *Rna*, 16(10):1870–1880, 2010.

[7] Robert M. Dirks and Niles A. Pierce. An algorithm for computing nucleic acid base-pairing probabilities including pseudoknots. *Journal of Computational Chemistry*, 25(10):1295–1304, 2004.

[8] Audrey Legendre, Eric Angel, and Fariza Tahi. Bi-objective integer programming for RNA secondary structure prediction with pseudoknots. *BMC Bioinformatics*, 19:13, January 2018.

[9] Patrick Gendron, Sébastien Lemieux, and François Major. Quantitative analysis of nucleic acid three-dimensional structures. *Journal of molecular biology*, 308(5):919–936, 2001.

[10] Michael Sarver, Craig L. Zirbel, Jesse Stombaugh, Ali Mokdad, and Neocles B. Leontis. FR3d: finding local and composite recurrent structural motifs in RNA 3d structures. *Journal of Mathematical Biology*, 56(1):215–252, January 2008.

[11] Xiang-Jun Lu, Harmen J. Bussemaker, and Wilma K. Olson. DSSR: an integrated software tool for dissecting the spatial structure of RNA. *Nucleic Acids Research*, 43(21):e142–e142, December 2015.

[12] Mahassine Djelloul and Alain Denise. Automated motif extraction and classification in RNA tertiary structures. *RNA*, 14(12):2489–2497, January 2008.

[13] Anton I. Petrov, Craig L. Zirbel, and Neocles B. Leontis. Automated classification of RNA 3d motifs and the RNA 3d Motif Atlas. *RNA*, 19(10):1327–1340, January 2013.

[14] Vladimir Reinharz, Antoine Soulé, Eric Westhof, Jérôme Waldispühl, and Alain Denise. Mining for recurrent long-range interactions in RNA structures reveals embedded hierarchies in network families. *Nucleic Acids Research*, 46(8):3841–3851, 2018.

[15] Grzegorz Chojnowski, Tomasz Waleń, and Janusz M Bujnicki. Rna bricks—a database of RNA 3d motifs and their interactions. *Nucleic acids research*, 42(D1):D123–D131, 2014.

[16] Cuncong Zhong, Haixu Tang, and Shaojie Zhang. RNAMotifScan: automatic identification of RNA structural motifs using secondary structural alignment. *Nucleic Acids Research*, 38(18):e176–e176, October 2010.

[17] Craig L. Zirbel, James Roll, Blake A. Sweeney, Anton I. Petrov, Meg Pirrung, and Neocles B. Leontis. Identifying novel sequence variants of RNA 3d motifs. *Nucleic Acids Research*, 43(15):7504–7520, September 2015.

[18] Roman Sarrazin-Gendron, Vladimir Reinharz, Carlos G Oliver, Nicolas Moitessier, and Jérôme Waldispühl. Automated, customizable and efficient identification of 3d base pair modules with bayespairing. *Nucleic acids research*, 2019.

[19] José Almeida Cruz and Eric Westhof. Sequence-based identification of 3d structural modules in RNA with rmdetect. *Nature methods*, 8(6):513, 2011.

[20] Vladimir Reinharz, François Major, and Jérôme Waldispühl. Towards 3d structure prediction of large RNA molecules: an integer programming framework to insert local 3d motifs in RNA secondary structure. *Bioinformatics*, 28(12):i207–i214, June 2012.

[21] Marc Parisien and Francois Major. The mc-fold and mc-sym pipeline infers RNA structure from sequence data. *Nature*, 452(7183):51, 2008.

[22] Mirela Andronescu, Vera Bereg, Holger H Hoos, and Anne Condon. Rna strand: the RNA secondary structure and statistical analysis database. *BMC bioinformatics*, 9(1):340, 2008.

[23] Kengo Sato, Yuki Kato, Michiaki Hamada, Tatsuya Akutsu, and Kiyoshi Asai. IPknot: fast and accurate prediction of RNA secondary structures with pseudoknots using integer programming. *Bioinformatics*, 27(13):i85–i93, July 2011.

[24] IBM ILOG. CPLEX: CPLEX Optimizer (academic license). https://www.ibm.com/analytics/optimization-modeling-interfaces, 2018.

Publishable brief abstract dedicated to a potential act of the JOBIM congress

Adaptation to animal sources of *Salmonella enterica* subsp. *enterica* deciphered by Genome Wide Association Study and Gene Ontology Enrichment Analysis at the pangenomic scale

Meryl VILA NOVA[1], Kévin LA[1], Kévin DURIMEL[1], Arnaud FELTEN[1], Philippe BESSIERES[2], Michel-Yves MISTOU[1], Mahendra MARIADASSOU[2] and Nicolas RADOMSKI[1]

[1] French Agency for Food, Environmental and Occupational Health & Safety (Anses), Genome Analysis Modelling Risk (GAMeR), 14 rue Pierre et Marie Curie, 94700, Maisons-Alfort, France

[2] National Institute of Agricultural Research (INRA), Applied Mathematics and Computer Science from Genomes to the Environment (MaIAGE), allée de Vilvert, 78352, Jouy-en-Josas, France

Corresponding Author: meryl.vilanova@anses.fr

Abstract

Salmonella enterica subsp. *enterica* is a public health issue related to food safety, and its adaptation to animal sources remains poorly described at the pangenome scale. Genome Wide Association Study (GWAS) from human genetics has recently been successfully adapted to bacteria to decipher the genomic determinants of host speciation, antibiotic resistance and virulence. In this study focusing on *Salmonella*, the combination of GWAS and Gene Ontology Enrichment Analysis (GOEA) will allow identification of genomic and metabolic signatures associated with animal sources.

As a first step, *Salmonella* mono- and multi-animal serovars were selected from a curated and synthesised subset of Enterobase, and the corresponding sequencing reads were downloaded from the European Nucleotide Archive (ENA) (i). Secondly, the accessory genes were detected from a pangenome performed with Roary based on assemblies produced with ARTWork-light. The coregenome variants (single nucleotide polymorphisms (SNPs) and small insertions/deletions (InDels)) were detected with the variant caller HaplotypeCaller implemented in iVARCall2 (ii). Taking into account variants from homologous recombination events, the accessory genes and coregenome variants were associated with animal sources using a microbial GWAS integrating an advanced correction of population structure implemented in GEMMA and (iii). Dependently of a local Uniprot dataset of GO-terms,, a GOEA was applied to emphasize metabolic pathways mainly impacted by the pangenomic mutations associated to animal sources (iv).

Based on the curated and synthesized subset of Enterobase, we established a dataset of 440 paired-end sequencing reads, including 15 serovars with associated spectra of animal sources, from generalist (i.e. multi-animal sources) to highly preferential (i.e. mono-animal sources) (i). In total, 19 130 accessory genes were listed, as well as 178 351 coregenome SNPs and InDels (ii). Among these mutations, 52 genomic signatures (iii) and 9 over-enriched metabolic signatures (iv) were associated to avian, bovine, swine and fish sources by GWAS and GOEA, respectively.

Our main conclusions show that genomic and metabolic determinants of *Salmonella* adaptation to animal sources may have been driven by the natural environment of the animal, specific physiological properties of the animal itself, environmental stimuli, distinct livestock diets, and work habits for health protection of livestock. We developed an integrated approach to screen pangenomic signatures of *Salmonella enterica* subsp. *enterica* associated with animal sources. The combination of a statically supported dataset of *Salmonella* genomes, a GWAS implementing an advanced population structure correction, and a GOEA integrating the most recent parent-child approach, allowed detection of mutations and metabolic pathways associated with animal sources.

Confidential and unpublishable detailed abstract dedicated to the scientific committee for the JOBIM congress

Adaptation to animal sources of *Salmonella enterica* subsp. *enterica* deciphered by Genome Wide Association Study and Gene Ontology Enrichment Analysis at the pangenomic scale

Meryl VILA NOVA[1], Kévin LA[1], Kévin DURIMEL[1], Arnaud FELTEN[1], Philippe BESSIERES[2], Michel-Yves MISTOU[1], Mahendra MARIADASSOU[2] and Nicolas RADOMSKI[1]

[1] French Agency for Food, Environmental and Occupational Health & Safety (Anses), Genome Analysis Modelling Risk (GAMeR), 14 rue Pierre et Marie Curie, 94700, Maisons-Alfort, France

[2] National Institute of Agricultural Research (INRA), Applied Mathematics and Computer Science from Genomes to the Environment (MaIAGE), allée de Vilvert, 78352, Jouy-en-Josas, France

Corresponding Author: meryl.vilanova@anses.fr

Abstract

Salmonella enterica subsp. enterica is a public health issue related to food safety, and its adaptation to animal sources remains poorly described at the pangenome scale. Genome Wide Association Study (GWAS) from human genetics has recently been successfully adapted to bacteria to decipher the genomic determinants of host speciation, antibiotic resistance and virulence. In this study focusing on Salmonella, the combination of GWAS and Gene Ontology Enrichment Analysis (GOEA) will allow identification of genomic and metabolic signatures associated with animal sources. As a first step, Salmonella serovars were selected from a curated and synthesised subset of Enterobase, and the corresponding sequencing reads were downloaded from the European Nucleotide Archive (ENA) (i). Secondly, the accessory genes and coregenome variants (single nucleotide polymorphisms (SNPs) and small insertions/deletions (InDels)) were detected (ii). Thirdly, the accessory genes and coregenome variants were associated to animal sources based on a GWAS integrating an advanced correction of the population structure (iii). Lastly, a GOEA was applied to emphasize metabolic pathways mainly impacted by the pangenomic mutations associated to animal sources (iv). Based on the curated and synthesized subset of Enterobase, we established a dataset of 440 paired-end sequencing reads and representing 15 serovars with associated spectra of animal sources, from generalist (i.e. multi-animal sources) to highly preferential (i.e. mono-animal sources) (i). In total, 19 130 accessory genes were listed by pangenome extraction applied on assembled genomes and 178 351 coregenome SNPs and InDels were detected by variant calling (ii). Among these mutations, 52 genomic signatures (iii) and 9 over-enriched metabolic signatures (iv) were associated to avian, bovine, swine and fish sources by GWAS and GOEA, respectively. Our main conclusions emphasise that genomic and metabolic determinants of Salmonella adaptation to animal sources may have been driven by the natural environment of the animal, specific physiological properties of the animal itself, environmental stimuli, distinct livestock diets, and work habits for health protection of livestock. We developed an integrated approach to screen pangenomic signatures of Salmonella enterica subsp. enterica associated to animal sources. The combination of a statically supported dataset of Salmonella genomes, a GWAS implementing an advanced population structure correction, and a GOEA integrating the most recent parent-child approach, allowed detection of mutations and metabolic pathways associated with animal sources.

Keywords

Food safety
Salmonella
Genome Wide Association Study
Gene Ontology Enrichment Analysis

1 Introduction

Salmonella is one of the main foodborne bacteria involved in human infections. In particular, serovars of *Salmonella enterica* subsp. *enterica* are responsible for about 80 million foodborne cases of gastroenteritis in developed countries per year [1,2]. Bacterial evolution is governed by stochastic point mutations due to replication errors, DNA damage, small insertions or deletions, as well as horizontal gene transfer promoted by recombination events [3]. Molecular biology highlighted that *Salmonella enterica* subsp. *enterica* extended over a wide range of host (birds, fish, reptiles, cattle ...) [4]. The millions of years between *Salmonella* and its common ancestor with *Escherichia coli* have led to the acquisition of genes involved in intestinal infection or colonization of tissues [5]. Without exhaustive and comprehensive data, some serovars are considered as specific to single hosts, while others are more associated with multiple hosts [6]. The mono-host specialisation of *Salmonella* serovars is frequently associated with specific pathologies (e.g. typhoid, paratyphoid, abortion, bacteraemia), while the generalist serovars are responsible of gastroenteritis [7]. Molecular mechanisms related to pathogenicity are encoded in *Salmonella* Pathogenic Islands (SPIs) and involved in invasion, survival and extra-intestinal propagation [8]. Despite the fact that the host adaptation of *Salmonella* serovars is poorly described at the pangenomic scale, some studies have demonstrated the role of SPIs in adaptation to avian [9] and bovine [10]. At the coregenome scale, the host adaptation of *Salmonella* serovars was explained by several fixed variants related to glutamate pathways [11]. The GWAS identify genomic variations associated with phenotypic traits of interest [12]. Over the last ten years, microbial GWAS has been applied to study persistence, host preference or virulence [13]. In contrast to human GWAS, microbial GWAS has to take into account confounding factors related to genome selection, homologous recombination events, population structure and multiple assays [14]. Despite its improvement since the beginning of the 21st century, the GOEA is rarely applied to bacterial genomes. The GOEA is used to test the hypergeometric distributions of the GO terms from a list of interest compared to a larger list, respecting the dependency between GO-terms (parent child approach) [15]. In a context of source tracking for food safety, the GOEA was applied to identify over-enriched metabolic pathways [16] among genes and variants associated by GWAS to animal sources.

2 Materials and Methods

We selected a dataset of 440 genomes of *Salmonella enterica* subsp. *enterica* from Enterobase (i). Then, we identified accessory genes and coregenome variants (SNPs and Indels) (ii). Thirdly, we associated these mutations with animal sources using a microbial GWAS implementing an advanced population structure correction (iii). Finally, we performed a GOEA to detect metabolic pathways mainly impacted by mutations associated with animal sources (iv).

2.1 Selection of a genome dataset (i)

We selected 440 samples according to data available from Enterobase. Respecting a high level of genomic diversity, we selected serovars from mono- and multi-animal serovars. The samples from environment, composite foods of the retail market and humans, were not retained because they are considered as vectors of pathogen expositions and exposed susceptible consumers in the present study. Based on a curated and synthesized subset of Enterobase, we selected 20 genomes of each of 3 serovars from potential mono-animal sources (avian, bovine, swine and fish) and between 60 and 80 genomes of each of 3 serovars from potential multi-animal sources.

2.2 Accessory genome (ii)

Based on assemblies produced with ARTWork-light, the pangenome was construct with Roary [17] setting 95% of identity for blastp and a strict definition of the coregenome.

2.3 Coregenome variants (ii)

The SNPs and InDels of the coregenome were detected with the variant caller HaplotypeCaller implemented in the iVARCall2 workflow [11], using *Salmonella* Typhimurium LT2 as a reference genome.

2.4 Genome Wide Association Study (iii)

We develop a microbial GWAS based on GEMMA [18], comparing different sizes of genome dataset, taking into account variants from homologous recombination events and checking population structure corrections.

The scripts can be found in the following GitHub repository: https://github.com/VilaNovaMeryl/microbialGWAS.

2.5 Gene Ontology Enrichment Analysis (iv)

We improved a workflow called fastGSEA based on a recently published version [11]. This workflow produces a fast GOEA dependently of a local Uniprot dataset of GO-terms. The scripts can be found in the following GitHub repository https://github.com/KDurimel/DNAlogy/tree/master/FAST_GOEA .

3 Results

3.1 Distributions of serovars from potential mono-and multi-animal sources

Respecting high levels of diversity in terms of phylogenomic relationships in view of previous studies [19], geographical origins, dates of isolation and BioProject accession numbers, a balanced dataset of *Salmonella* serovars potentially from mono- and multi-animal sources (figure 1) were selected in order to detect mutations and related metabolic pathways associated to animal sources.

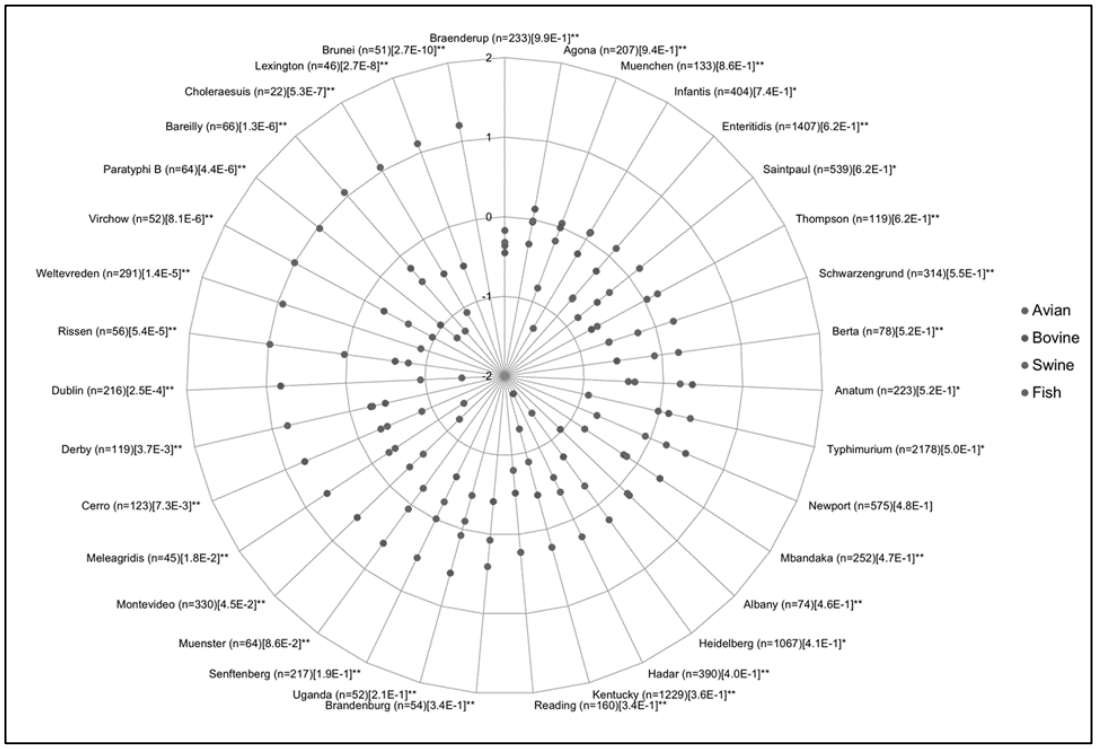


Fig 1. Common logarithmic values of relative proportions of serovars of *Salmonella enterica* subsp. *enterica* per animal source corrected by animal source distribution observed in a curated and synthesized subset of Enterobase retaining samples fully described and linked to corresponding reads from BioProject.

3.2 Phylogenomic relationships between serovars from potential mono-and multi-animal sources

With the exception of the polyphyletic serovars Newport and Cerro, all the genomes of the others serovars were clustered together (figure 2) based on three phylogenomic reconstructions. This coexistence of purely clonal and nearly panmictic (i.e. multi-animal sources) serovars (figure 2), emphases the necessity to correct the population structure before to associate mutations to animal sources by microbial GWAS.

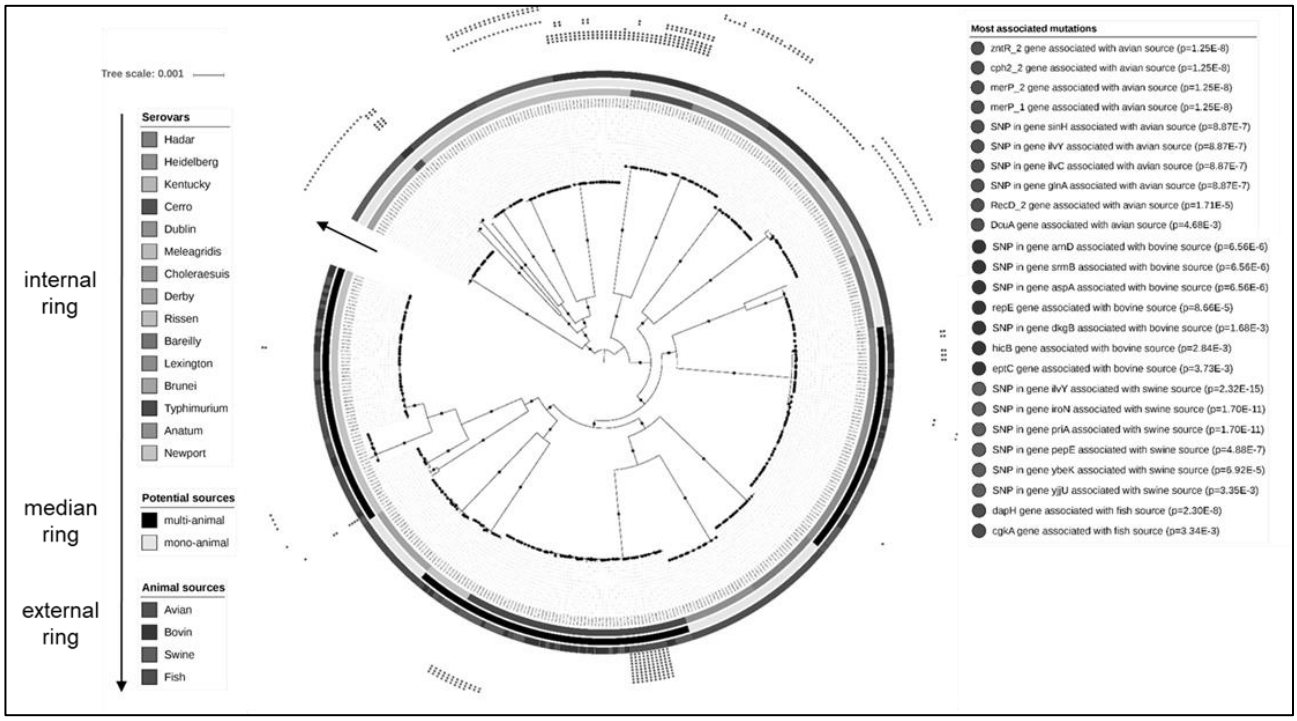


Fig 2. Phylogenomic inference by maximum likelihood aiming to reconstruct population structure of genomes of *Salmonella enterica* subsp. *enterica* serovars (n=440) from potential mono- and multi-animal sources

3.3 Mutation associated with animal sources (i.e. microbial GWAS)

In view to higher functional impacts of accessory genes compared to coregenome variants, 38 genes were detected as associated with animal sources, whereas only 3 intergenic, 3 synonymous and 8 non-synonymous variants were associated to these traits of interest (table 1). Because synonymous variants associated to traits of interest (table 1) may emphasize elements of regulation [20] or phenotypical impacts [21], we decided to retain them to perform GOEA.

Mutations	Annotations			Before GWAS		After GWAS			
				Including homologous recombination	Excluding homologous recombination	Avian source	Bovine source	Swine source	Fish source
accessory genes and variants	annotated, hypothetical and intergenic			178 351	38 837	18	16	11	7
accessory genes		annotated		6 387	6 387	6	3	0	2
		hypothetical		12 743	12 743	8	9	5	5
coregenome variants	intergenic			17 362	2 288	1	1	1	0
	intragenic	synonymous		68 157	8 365	1	1	1	0
		non synonymous	missenses	65 044	8 017	2	2	4	0
			start lost	144	19	0	0	0	0
			stop gained	4 202	525	0	0	0	0
			frameshift	1 019	136	0	0	0	0
			disruptive inframe insertions	122	14	0	0	0	0
			disruptive inframe deletions	204	31	0	0	0	0
	multiple annotations			2 967	312	0	0	0	0

Tab 1. Mutations before and after microbial GWAS aiming to associate animal sources with mutations from accessory and coregenome of *Salmonella enterica* subsp. *enterica* serovars (n=440)

3.4 Metabolic pathways mainly impacted by mutations associated with animal sources (i.e. GOEA)

3.5 Concisely 6, 1, 0 and 2 GO-terms of interest were retrieve concerning the avian, bovine, swine and fish sources, respectively (table 2). All studied phenotypic traits included; these GO-terms of interest were related to rare GO-terms in comparison with the pangenomic GO-terms (table 2). These GO-terms of interest were mainly related to molecular functions (66%) and biological processes (33%).

Animal source	Uniprotkb	Associated Mutations	GO-term identifier	GO-term	Hits	Exp. hits	GO level	Corr. p-value	Ontology
avian	Q55434	gene *cph2_2*	GO:0009585	red, far-red light phototransduction	1	0.01	7	1E-07	BP
avian	Q55434	gene *cph2_2*	GO:0009584	detection of visible light	1	0.01	7	1E-07	BP
avian	Q55434	gene *cph2_2*	GO:0009883	red or far-red light photoreceptor activity	1	0.01	5	1E-07	MF
avian	Q9RT63	gene *recD2*	GO:0043141	ATP-dependent 5'-3' DNA helicase activity	1	0.01	11	1E-07	MF
avian	Q9RT63	gene *recD2*	GO:0008094	DNA-dependent ATPase activity	5	0.28	10	1E-03	MF
avian	P0ABN5	gene *dcuA*	GO:0015740	C4-dicarboxylate transport	3	0.13	10	2E-02	BP
bovine	Q7CPA1	SNP in *aspA*	GO:0008797	aspartate ammonia-lyase activity	1	0.01	6	1E-07	MF
fish	Q7A2S0	gene *dapH*	GO:0047200	tetrahydrodipicolinate N-acetyltransferase activity	1	0.01	8	1E-07	MF
fish	P43478	gene *cgkA*	GO:0033918	kappa-carrageenase activity	1	0.01	6	1E-07	MF

Tab 2. GO-terms mainly enriched by GOEA applied on accessory genes and coregenome variants of *Salmonella enterica* subsp. *enterica* serovars associated by microbial GWAS with animal sources

4 Discussion

Based on a balanced and diverse dataset of genomes, we have been able to identify genomic signatures associated with animal sources. Signals of host adaptation were previously identified in *Staphylococcus aureus* [23] and *Campylobacter* [24]. The mutations identified in the present study could be used as *in silico* or *in vitro* markers in the context of source monitoring related to food safety [25]. Following the same distribution within the avian genomes, the associated mutations belong to genes *zntR2*, *cph2-2*, *merP-1* and *merP-2*. The gene *zntR2* allows binding to DNA-related sites [26]. The gene *cph2-2* activates the mobility capacity to red light and may be related to poultry growth conditions [27]. The periplasmic components of the mercuric transport protein may be a sign of adaptation to mercury exposure. The exposure to the contaminated biosphere [29] and/or mercury-based vaccines [30] may be the origins of this adaptation. The mutations associated with bovine sources affected the metabolic process related to the activity of aspartate ammonia-lyase. The corresponded gene *aspA* converts aspartate into fumarate which is reduced to succinate [31]. This process was also observed in *Escherichia coli* and could promote the adaptation of *Salmonella* to the bovine intestine [32]. A SNP associated with pig sources affected the dideptidase E, which is involved in the sequestration of the aspartate peptide in the synthesis of aspartate amino acids [33]. Another associated SNP affected the primosomal N protein, which allows restarting of blocked replication forks via its helicase activity [34]. Regarding fish sources, the associated mutations affected the metabolic processes involved in the activities of kappa-carrageenase and tetrahydrodipicolinate. One participates in the degradation of a sulphated linear polysaccharide (k-carrageenan) [35], while the other is known as the first step in the biosynthesis of L-lysine [36].

Acknowledgements

The French Agency for Food, Environmental and Occupational Health and Safety (Anses) and The French National Institute for Agricultural Research (INRA) (grant name Typautobac) supported this work. We thank especially Pierre-Yves Letournel and Thomas Texier for providing high-performance computing resources.

References

[1] Majowicz SE, Musto J, Scallan E, Angulo FJ, Kirk M, O'Brien SJ, Jones TF, Fazil A, Hoekstra RM. 2010. The Global Burden of Nontyphoidal *Salmonella* Gastroenteritis. Clin. Infect. Dis. 50:882–889.

[2] EFSA-ECDC. 2016. The European Union summary report on trends and sources of zoonoses, zoonotic agents and food-borne outbreaks in 2015. Eur. Food Saf. Auth. J. 14.

[3] Sheppard SK, Guttman DS, Fitzgerald JR. 2018. Population genomics of bacterial host adaptation. Nat. Rev. Genet. 19:549–565.

[4] Evangelopoulou G, Kritas S, Govaris A, Burriel AR. 2013. Animal salmonelloses: a brief review of "host adaptation and host specificity" of Salmonella spp. Vet. World 6:703–708.
[5] Bäumler AJ, Tsolis RM, Ficht TA, Adams LG. 1998. Evolution of host adaptation in Salmonella enterica. Infect. Immun. 66:4579–4587.
[6] Porwollik S, Santiviago CA, Cheng P, Florea L, McClelland M. 2005. Differences in Gene Content between Salmonella enterica Serovar Enteritidis Isolates and Comparison to Closely Related Serovars Gallinarum and Dublin. J. Bacteriol. 187:6545–6555.
[7] Tanner JR, Kingsley RA. 2018. Evolution of Salmonella within Hosts. Trends Microbiol. 26:986–998.
[8] Siriken B. 2013. Salmonella Pathogenicity Islands. Mikrobiyol. Bul.:181–188.
[9] Thomson NR, Clayton DJ, Windhorst D, Vernikos G, Davidson S, Churcher C, Quail MA, Stevens M, Jones MA, Watson M, et al. 2008. Comparative genome analysis of Salmonella Enteritidis PT4 and Salmonella Gallinarum 287/91 provides insights into evolutionary and host adaptation pathways. Genome Res. 18:1624–1637.
[10] Langridge GC, Fookes M, Connor TR, Feltwell T, Feasey N, Parsons BN, Seth-Smith HMB, Barquist L, Stedman A, Humphrey T, et al. 2015. Patterns of genome evolution that have accompanied host adaptation in *Salmonella*. Proc. Natl. Acad. Sci. 112:863–868.
[11] Felten A, Vila Nova M, Durimel K, Guillier L, Mistou M-Y, Radomski N. 2017. First gene-ontology enrichment analysis based on bacterial coregenome variants: insights into adaptations of Salmonella serovars to mammalian- and avian-hosts. BMC Microbiol. 17:1–22.
[12] Lees JA, Bentley SD. 2016. Bacterial GWAS: not just gilding the lily. Nat. Rev. Microbiol. 14:406–406.
[13] Bartha I, Carlson JM, Brumme CJ, McLaren PJ, Brumme ZL, John M, Haas DW, Martinez-Picado J, Dalmau J, López-Galíndez C, et al. 2013. A genome-to-genome analysis of associations between human genetic variation, HIV-1 sequence diversity, and viral control. eLife 2.
[14] Power RA, Parkhill J, de Oliveira T. 2017. Microbial genome-wide association studies: lessons from human GWAS. Nat. Rev. Genet. 18:41–50.
[15] Grossmann S, Bauer S, Robinson PN, Vingron M. 2007. Improved detection of overrepresentation of Gene-Ontology annotations with parent child analysis. Bioinformatics 23:3024–3031.
[16] Lee I-H, Lee K, Hsing M, Choe Y, Park J-H, Kim SH, Bohn JM, Neu MB, Hwang K-B, Green RC, et al. 2014. Prioritizing Disease-Linked Variants, Genes, and Pathways with an Interactive Whole-Genome Analysis Pipeline. Hum. Mutat. 35:537–547.
[17] Page AJ, Cummins CA, Hunt M, Wong VK, Reuter S, Holden MTG, Fookes M, Falush D, Keane JA, Parkhill J. 2015. Roary: rapid large-scale prokaryote pan genome analysis. Bioinformatics 31:3691–3693.
[18] Earle SG, Wu C-H, Charlesworth J, Stoesser N, Gordon NC, Walker TM, Spencer CCA, Iqbal Z, Clifton DA, Hopkins KL, et al. 2016. Identifying lineage effects when controlling for population structure improves power in bacterial association studies. Nat. Microbiol. 1:16041.
[19] Roer L, Hendriksen RS, Leekitcharoenphon P, Lukjancenko O, Kaas RS, Hasman H, Aarestrup FM. 2016. Is the Evolution of *Salmonella enterica* subsp. *enterica* Linked to Restriction-Modification Systems?Eisen J, editor. mSystems 1.
[20] Shabalina SA, Spiridonov NA, Kashina A. 2013. Sounds of silence: synonymous nucleotides as a key to biological regulation and complexity. Nucleic Acids Res. 41:2073–2094.
[21] Hammarlöf DL, Kröger C, Owen SV, Canals R, Lacharme-Lora L, Wenner N, Schager AE, Wells TJ, Henderson IR, Wigley P, et al. 2018. Role of a single noncoding nucleotide in the evolution of an epidemic African clade of *Salmonella*. Proc. Natl. Acad. Sci. 115:E2614–E2623.
[22] Pascoe B, Méric G, Yahara K, Wimalarathna H, Murray S, Hitchings MD, Sproston EL, Carrillo CD, Taboada EN, Cooper KK, et al. 2017. Local genes for local bacteria: Evidence of allopatry in the genomes of transatlantic *Campylobacter* populations. Mol. Ecol. 26:4497–4508.
[23] Lowder BV, Guinane CM, Ben Zakour NL, Weinert LA, Conway-Morris A, Cartwright RA, Simpson AJ, Rambaut A, Nubel U, Fitzgerald JR. 2009. Recent human-to-poultry host jump, adaptation, and pandemic spread of Staphylococcus aureus. Proc. Natl. Acad. Sci. 106:19545–19550.
[24] Sheppard SK, Colles F, Richardson J, Cody AJ, Elson R, Lawson A, Brick G, Meldrum R, Little CL, Owen RJ, et al. 2010. Host Association of Campylobacter Genotypes Transcends Geographic Variation. Appl. Environ. Microbiol. 76:5269–5277.
[25] Wheeler NE. 2019. Tracing outbreaks with machine learning. Nat. Rev. Microbiol. [Internet]. Available from: http://www.nature.com/articles/s41579-019-0153-1
[26] Rodriguez-Maillard JM, Arutyunov D, Frost LS. 2010. The F plasmid transfer activator TraJ is a dimeric helix-turn-helix DNA-binding protein: F TraJ binds DNA in vivo. FEMS Microbiol. Lett. 310:112–119.
[27] Prayitno D, Phillips C, Omed H. 1997. The effects of color of lighting on the behavior and production of meat chickens. Poult. Sci. 76:452–457.
[28] Carravieri A, Cherel Y, Blévin P, Brault-Favrou M, Chastel O, Bustamante P. 2014. Mercury exposure in a large subantarctic avian community. Environ. Pollut. 190:51–57.
[29] Lombardi G, Lanzirotti A, Qualls C, Socola F, Ali A-M, Appenzeller O. 2012. Five Hundred Years of Mercury Exposure and Adaptation. J. Biomed. Biotechnol. 2012:1–10.
[30] Stone HD. 1985. Effect of thimerosal concentration on the efficacy of inactivated Newcastle disease oil-emulsion vaccines. Avian Dis. 29:1030-1035.

[31] Lacey M, Agasing A, Lowry R, Green J. 2013. Identification of the YfgF MASE1 domain as a modulator of bacterial responses to aspartate. Open Biol. 3:130046–130046.
[32] Bertin Y, Segura A, Jubelin G, Dunière L, Durand A, Forano E. 2018. Aspartate metabolism is involved in the maintenance of enterohaemorrhagic *Escherichia coli* O157:H7 in bovine intestinal content. Environ. Microbiol. 20:4473–4485.
[33] Conlin CA, Håkensson K, Liljas A, Miller CG. 1994. Cloning and nucleotide sequence of the cyclic AMP receptor protein-regulated Salmonella typhimurium pepE gene and crystallization of its product, an alpha-aspartyl dipeptidase. J. Bacteriol. 176:166–172.
[34] Manhart CM, McHenry CS. 2013. The PriA Replication Restart Protein Blocks Replicase Access Prior to Helicase Assembly and Directs Template Specificity through Its ATPase Activity. J. Biol. Chem. 288:3989–3999.
[35] Manuhara GJ, Praseptiangga D, Riyanto RA. 2016. Extraction and Characterization of Refined K-carrageenan of Red Algae [Kappaphycus Alvarezii (Doty ex P.C. Silva, 1996)] Originated from Karimun Jawa Islands. Aquat. Procedia 7:106–111.
[36] Simms SA, Voige WH, Gilvarg C. 1984. Purification and characterization of succinyl-CoA: tetrahydrodipicolinate N-succinyltransferase from Escherichia coli. J. Biol. Chem. 259:2734–2741.

Allele-specific analysis of epigenetic and transcriptomic data to study Drosophila developmental *cis*-regulatory architecture.

Swann Floc'hlay[1], Bingqing Zhao[2], David Garfield[3], Emily Wong[4], Raquel Marco-Ferreres[5], Morgane Thomas-Chollier[1], Denis Thieffry[1] and Eileen Furlong[5]

[1] Institute of Biology of the Ecole normale supérieure, Computational Systems Biology laboratory, 46 rue d'ULM, 75005, Paris, France.
[2] Stanford University, Snyder laboratory, 300 Pasteur drive, 94305, Stanford, USA.
[3] IRI for the Life Sciences, Garfield laboratory for Evolutionary Biology, Philippstr.13, 10115, Berlin, Germany.
[4] School of Biomedical Sciences, Queensland University, QLD 4072, Brisbane, Australia.
[5] European Molecular Biology Laboratory, Furlong laboratory, Meyerhofstraße 1, 69117, Heidelberg, Germany.

Corresponding author: furlong@embl.de, denis.thieffry@ens.fr

Abstract *Recent high-throughput sequencing studies between individuals of a given species have revealed extensive variation in gene expression, as a consequence of segregating genetic variation within the population. Most of this regulatory genetic variation is in non-coding DNA, presumably disrupting the function of enhancer elements. However, understanding and predicting how genetic variants disrupt transcriptional regulation remains very poorly understood.*

We aim at getting a mechanistic understanding of how natural genetic variation affects multiple layers of transcriptional regulation. We use hybrid embryos of genetically distinct Drosophila lines, isolated from a wild population, at three crucial time windows of embryonic development. The use of hybrid individuals offers a powerful approach to dissect cis versus trans-regulatory mutations by obtaining allele specific information (e.g. allelic specific ATAC-seq, ChIP-seq, RNA-seq data).

This analysis offers some interesting bioinformatic challenges, such as mapping biases in genetically distinct lines and confounding factors in correlation analysis. To control for these effects, we have used the parental genome mapping strategy and the partial correlation method. Surprisingly enough, the regulatory architecture obtained by the allelic ratio correlation analysis differs noticeably from results obtained solely from coverage.

The integration of these various levels of regulation should lead to a more extensive view of the genetic bases influencing transcriptional regulation by simultaneously integrating data on gene expression, enhancer/promoter activity and chromatin states.

Keywords Allele-specific analysis, *Drosophila melanogaster*, transcriptomics, epigenomics, partial correlation analysis.

1 Introduction

According to various GWAS (Genome Wide Association) studies, the vast majority of Single Nucleotide Polymorphisms (SNPs) associated with genetic diseases lays within the non-coding regions of the genome. They presumably disrupt regulatory elements, such as enhancers and promoters. Despite these evidences for a link between genetic variations and phenotypes, the mechanisms causing these regulatory changes are not yet clearly understood [1,2]. This study aims at understanding the impact of natural genetic variation on transcriptional regulation.

Based on a comprehensive dataset generated by a series of high-throughput experiments in the Furlong laboratory (EMBL, Heidelberg), we have a chance to look more deeply at the interplays between the various layers of transcriptional regulation, spanning from changes in chromatin accessibility (ATAC-seq) and epigenetic remodelling (ChIP-seq histone) to variations in gene expression (RNA-seq). These experiments have been realised at three developmental stages, in eight different crosses of paternal *Drosophila* lines from the DGRP collection [3] with a common maternal line (Fig. 1). Such

experimental design offers the possibility to dissect the regulatory changes in the context of genetic variations, using allele-specific measurements.

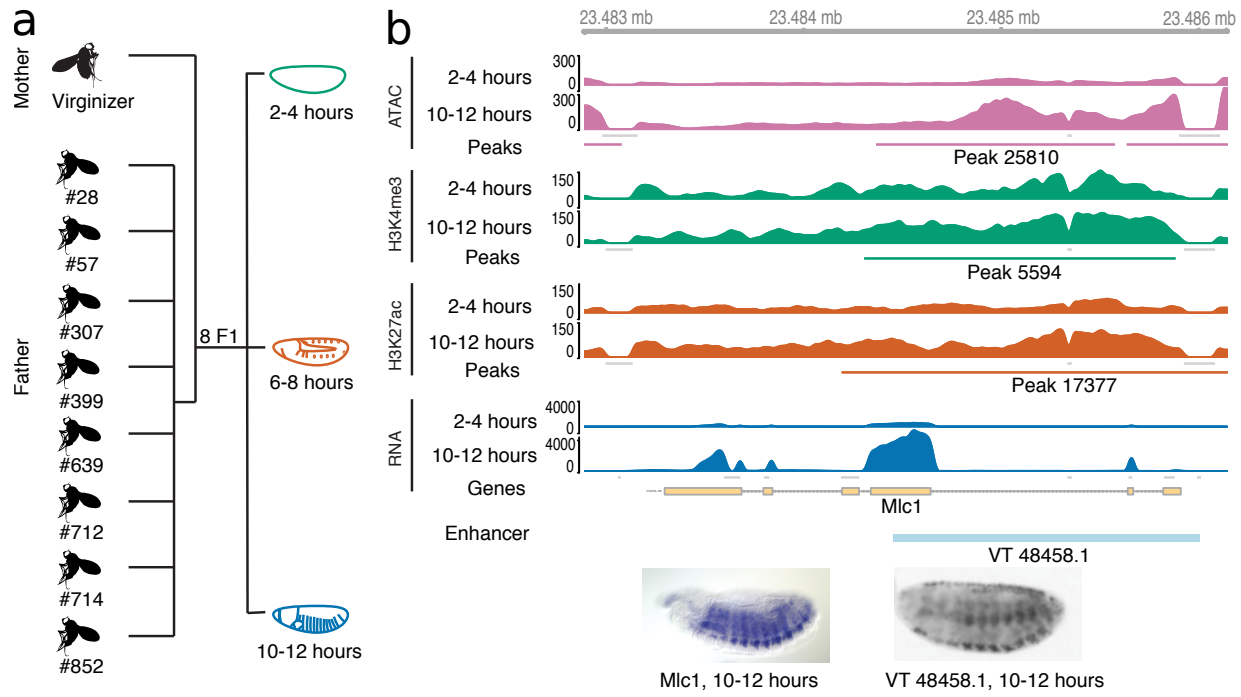


Fig. 1. a. Summary of the experimental design. F1 embryos have been sampled from 8 crosses of isogenic fly lines. Our measurements were made at three developmental time points: 2-4 hours after egg laying, consisting primarily of pre-gastrulation embryos, 6-8 hours, during cell-fate specification, and 10-12 hours, during terminal differentiation of the major tissue lineages. **b.** ATAC-seq, histone ChIP-seq on the epigenetic marks H3K27ac and H3K4me3, and RNA-seq profiling experiments have been performed (2 replicates per sample). Focus on the signal tracks obtained in the region of Mlc1 and its proximal enhancer. Courtesy of B. Zhao for the *in-situ* pictures.

2 Data processing

The first challenge of this study consisted in building a computational pipeline to map the sequences generated by the high-throughput experiments onto reference genomes. Our pipeline is using the personalised genome strategy, in order to prevent mappability issues related to the genetic diversity of our samples [4]. In addition, we built a pipeline for the processing of simulated transcriptomic and genomic reads, to blacklist the regions showing a coverage bias in the mapping process.

The F1 lines come from isogenic crosses, each involving a different paternal line. The sequenced embryo genomes are therefore highly heterozygous. Using this feature, we can assign each read properly mapped to its parent of origin, based on its sequence and on the available information on SNPs segregating in each cross. This way, we can construct parent-specific count tables for each sample, and more interestingly, measure the allele-specific expression (ASE) imbalance for each feature (transcript or peak regions). Evidence for imbalance is then tested for each feature by fitting the parent-specific counts to a beta-binomial distribution with the dispersion parameter fitted for each type of data.

Based on genomic DNA data and RNA-seq data of unfertilized eggs, we have successfully removed additional biases contributing to a shift in the ASE distribution. This way, we have discarded from the analysis (i) genes with evidence of maternal transcript deposition and (ii) SNPs showing evidence of genotyping errors or inherent imbalance.

3 Analysis

We characterised the ASE distribution and signal quality of our data. Moreover, as these results span three dimensions (namely time, genetic background and transcriptional regulatory layer), the next challenge is to investigate the ASE dynamics and correlations across these variables. In order to discriminate the signal coming from enhancers and promoters, we segregated the overlaps between the

four different regulatory layers into a TSS proximal (+/- 500bp) and a TSS distal sets. As expected, almost all H3K4me3 peaks are present in the proximal set. Furthermore, overlapping signal from ATAC-seq and H3K27ac ChIP-seq occur at potential enhancer TSS-distal regions.

Correlations among data types provide insights into the ways in which *cis*-acting genetic variation influences gene regulatory phenotypes, but interpretation is confounded by the tight interdependencies between the four datatypes. To assess the independent influences of changes in histone modifications and chromatin accessibility on gene expression, we used a partial correlations method, a technique for identifying independent, pairwise correlations within datasets consisting of multiple, co-varying variables, which has been previously applied to similar ChIP-seq data in mammalian cell culture [5]. As applied to total count data (i.e. coverage data), our results are highly concordant with previously reported results, with a strong independent correlation between H3K27ac and RNA (Fig. 2a).

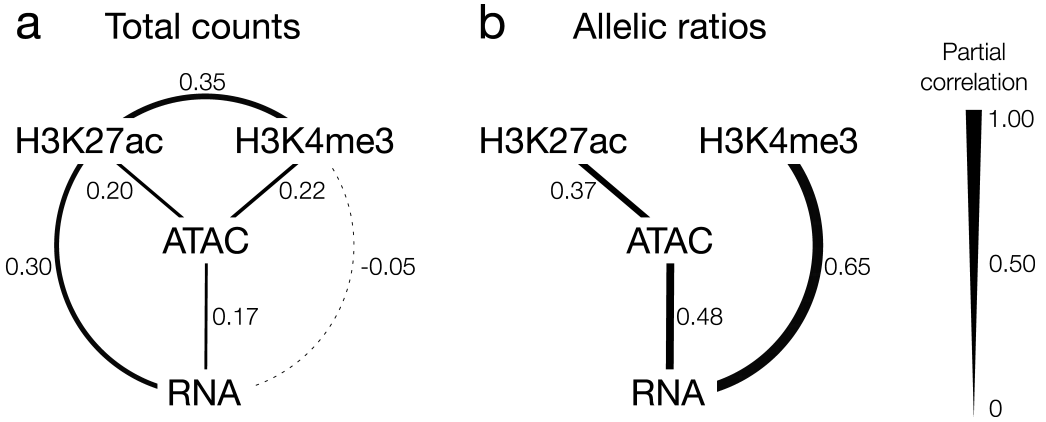


Fig. 2. Partial correlation results between TSS-proximal non-coding features and gene expression levels. Line width represent the strength of the partial correlation. Solid lines and dashed lines represent significantly positive and negative correlations respectively. Absence of line between two regulatory layers stand for a lack of significant correlation. **a.** Partial correlation on total count data. **b.** Partial correlation on allelic ratio data.

Correlations in allele-specific ratios, however, tell a different story. Interestingly, we see no evidence for a direct relationship between H3K27ac and RNA, but instead a direct correlation between allelic imbalance in H3K4me3 peaks and imbalance in the transcription of associated genes. Although we note overall higher correlation values in allelic imbalance analysis compared to the total counts, we also see a decrease in the number of significant direct relationships due to the loss of evidence for independent correlation between chromatin accessibility and H3K4me3, and the two histone modifications. The loss of a direct relationship between regulatory layers may be due to the presence of another confounding variable, which "explains-away" the observed indirect correlation. For instance, in the allelic ratio partial correlation results, we see no evidence for a direct relationship between chromatin accessibility and H3K4me3, probably because gene expression and H3K27ac allelic imbalance measures act as confounding factors explaining the largest fraction of variance within the correlation. Hence, at constant gene expression and H3K27ac allelic imbalance, we do not see any significant correlation between H3K4me3 and chromatin accessibility.

The contrasting importance of H3K4me3 and H3K27ac between allelic ratio and total count correlations could support the hypothesis that mutations that most impact gene expression allelic imbalance may not act via the pathways most closely correlated with total gene expression level. In order to further investigate this insight for a regulatory difference, we aim at deciphering the impact of the mutations on transcription factor binding affinities. Indeed, the detection of a given SNP co-segregating with the presence or absence of allelic imbalance in an underlying gene or non-coding features could be the hallmark of a mutation disrupting a transcription factor binding motif necessary for gene expression regulation. To detect such event, we will develop a pipeline able to detect enriched binding motifs in cis-regulatory regions, based on ATAC and ChIP histone signal. This work will be based on a pre-existing package, RSAT (rsat.eu) [6], and developments will be made freely accessible to the community and applicable to any organism.

The investigation of the impact of SNPs on transcription factors is an exciting area of research, which has already shown some conclusive results, such as the discovery of the pioneer role of the transcription factor Grainy head in epithelial specification [7].

Acknowledgements

This work was supported by the Furlong laboratory, the Computational Systems Biology team and the doctoral school ED515 from PSL Research University.

References

[1] Sierra S. Nishizaki and Alan P. Boyle. Mining the Unknown: Assigning Function to Noncoding Single Nucleotide Polymorphisms. *Trends in Genetics*, 33(1):34–45, January 2017.

[2] Eileen E. M. Furlong and Michael Levine. Developmental enhancers and chromosome topology. *Science*, 361(6409):1341–1345, September 2018.

[3] Trudy F. C. Mackay, Stephen Richards, Eric A. Stone, Antonio Barbadilla, Julien F. Ayroles, Dianhui Zhu, Sònia Casillas, Yi Han, Michael M. Magwire, Julie M. Cridland, Mark F. Richardson, Robert R. H. Anholt, Maite Barrón, Crystal Bess, Kerstin Petra Blankenburg, Mary Anna Carbone, David Castellano, Lesley Chaboub, Laura Duncan, Zeke Harris, Mehwish Javaid, Joy Christina Jayaseelan, Shalini N. Jhangiani, Katherine W. Jordan, Fremiet Lara, Faye Lawrence, Sandra L. Lee, Pablo Librado, Raquel S. Linheiro, Richard F. Lyman, Aaron J. Mackey, Mala Munidasa, Donna Marie Muzny, Lynne Nazareth, Irene Newsham, Lora Perales, Ling-Ling Pu, Carson Qu, Miquel Ràmia, Jeffrey G. Reid, Stephanie M. Rollmann, Julio Rozas, Nehad Saada, Lavanya Turlapati, Kim C. Worley, Yuan-Qing Wu, Akihiko Yamamoto, Yiming Zhu, Casey M. Bergman, Kevin R. Thornton, David Mittelman, and Richard A. Gibbs. The *Drosophila melanogaster* Genetic Reference Panel. *Nature*, 482(7384):173–178, February 2012.

[4] Kraig R. Stevenson, Joseph D. Coolon, and Patricia J. Wittkopp. Sources of bias in measures of allele-specific expression derived from RNA-seq data aligned to a single reference genome. *BMC Genomics*, 14(1):536, August 2013.

[5] Julia Lasserre, Ho-Ryun Chung, and Martin Vingron. Finding Associations among Histone Modifications Using Sparse Partial Correlation Networks. *PLOS Computational Biology*, 9(9):e1003168, September 2013.

[6] Nga Thi Thuy Nguyen, Bruno Contreras-Moreira, Jaime A. Castro-Mondragon, Walter Santana-Garcia, Raul Ossio, Carla Daniela Robles-Espinoza, Mathieu Bahin, Samuel Collombet, Pierre Vincens, Denis Thieffry, Jacques van Helden, Alejandra Medina-Rivera, and Morgane Thomas-Chollier. RSAT 2018: regulatory sequence analysis tools 20th anniversary. *Nucleic Acids Research*, 46(W1):W209–W214, July 2018.

[7] Jelle Jacobs, Mardelle Atkins, Kristofer Davie, Hana Imrichova, Lucia Romanelli, Valerie Christiaens, Gert Hulselmans, Delphine Potier, Jasper Wouters, Ibrahim I. Taskiran, Giulia Paciello, Carmen B. González-Blas, Duygu Koldere, Sara Aibar, Georg Halder, and Stein Aerts. The transcription factor Grainy head primes epithelial enhancers for spatiotemporal activation by displacing nucleosomes. *Nature Genetics*, 50(7):1011, July 2018.

CISPER: Computational Identification of Switch Points (in a Metabolic Network) within an Environmental Range

Francis Mairet
Ifremer, Physiology and Biotechnology of Algae laboratory, rue de l'Ile d'Yeu, 44311 Nantes, France

Corresponding author: francis.mairet@ifremer.fr

Abstract *A key challenge in systems biology is to identify, from the hundreds or thousands of molecules involved in a metabolic network, the key metabolites where the orientation of fluxes occurs. These switch nodes can be used to decompose the metabolic network into modules, to develop reduced dynamical models, to study cellular regulations, to reroute cellular fluxes (by controlling environmental conditions or by metabolic engineering)... Here, we propose a method - called CISPER - to identify these switch points based on the analysis of a set of flux balance analysis (FBA) solutions. A metabolite is considered as a switch if the fluxes at this point are redirected in a different way when conditions change. More precisely, the optimal solution for each condition is computed using FBA. Then, for each metabolite, we consider the flux vectors (including stoichiometry) of the reactions involving this metabolite (as a substrate or a product) for all the conditions. If the dimension of the vector space generated by this set is greater than one (i.e. the flux vectors involving this metabolite for different conditions are not co-linear), the metabolite is considered as a switch node. After its presentation, the soundness of CISPER is shown with two case studies: the central metabolism of the microalga Tisochrysis lutea and the transition from aerobic to anaerobic conditions in the yeast Saccharomyces cerevisiae.*

Keywords Metabolic Networks, Flux Balance Analysis, Branching point, Decomposition into subnetworks.

1 Introduction

Metabolic networks represent a key tool in systems biology. More and more networks are now available, with a more complete coverage of the metabolism. Flux Balance Analysis (FBA) uses these networks to predict all the metabolic fluxes [1]. This method is based on two main assumptions. First, the metabolism is considered at steady state (corresponding to balanced growth condition). Second, an objective function to be maximized is defined (e.g. the specific growth rate for microorganisms). The metabolic fluxes are then the solution of a linear optimization problem (also called LP problem, for Linear Programming), which can easily be solved numerically. Already very successfully used, FBA is nonetheless restricted to balanced growth. The analysis of metabolic network in dynamical conditions is more tricky, and several methods (such as DRUM [2]) propose to decompose the whole network into different modules to tackle this challenge.

Methods for network splitting have been proposed, based on network topology, flux coupling, or elementary flux mode (reviewed in [3]). Here, we adopt another viewpoint. We want to identify switch points, corresponding to key metabolites where the fluxes are redirected in a different way for a given set of conditions. The metabolic network can then be directly decomposed into subnetworks connecting the switch metabolites. Our method - called CISPER - is briefly described below and then two case studies are presented.

2 Method

CISPER is based on the analysis of a set of FBA solutions for a range of environmental conditions (e.g. different inputs, different objective functions reflecting different metabolic stages...). A metabolite is considered as a switch if the fluxes at this point are redirected in a different way when conditions change. This is illustrated in Figure 1. On the top, the distribution of fluxes occurs always in the same way (i.e. one third of the incoming flux R1 goes to R2, the remaining goes to R3), so x_1 is not a

switch point. On the contrary, in the bottom example, the incoming flux is rerouted according to the conditions, so x_1 is considered in this case as a switch point. This simple principle can be evaluated numerically using linear algebra.

More precisely, the optimal solution for each condition is computed using FBA. Then, for each metabolite, we consider the flux vectors (including stoichiometry) of the reactions involving this metabolite (as a substrate or a product) for all the conditions. If the dimension of the vector space generated by this set is greater than one (i.e. the flux vectors involving this metabolite for different conditions are not co-linear), the metabolite is considered as a switch node. This dimension is evaluated by singular value decomposition (SVD): if the second singular value is greater than a given tolerance, the metabolite is selected. Additionally, this singular value gives a score to represent the significance of each switch node. All the metabolites can then be ranked according to their score, and one can fix the cute-off value to select a given number of switches.

The CISPER method has been implemented under Matlab within the COBRA framework [4]. A toolbox is under development and will soon be available.

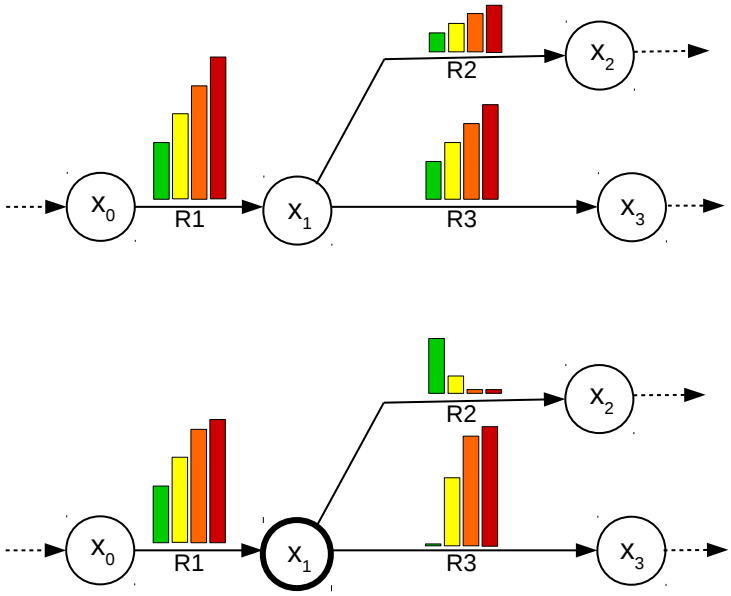


Fig. 1. Principe for switch point identification in metabolic networks. The circles x_i and the arrows R_j represent respectively metabolites and reactions. The colored bars show reaction fluxes (computed by Flux Balance Analysis) for different conditions (each color represents a condition). Top: the orientation of fluxes occurs always in the same way, so x_1 is not a switch point. Down: the incoming flux is rerouted according to conditions, so x_1 is a switch point.

3 Results

CISPER is used on two case studies: the central metabolism of the microalga *Tisochrysis lutea* and the transition from aerobic to anaerobic conditions in the yeast *Saccharomyces cerevisiae*. In both cases, biomass maximization was used as the objective function for FBA. The cute-off value for CISPER was adapted to select only just a few switch metabolites (without considering cofactors).

3.1 Carbon accumulation in the microalga *Tisochrysis lutea*

We consider the core metabolic network of the microalga *Tisochrysis lutea* [2], which is composed of 157 metabolites and 160 reactions. Microalgae are known to accumulate carbon storage (carbohydrate and neutral lipids) under nitrogen limitation [5]. To mimic such stress, the stoichiometry of the biomass synthesis reaction is changed, from low to high carbon accumulation. Using CISPER, we obtain the following switch nodes: glyceraldehyde 3-phosphate, acetyl-CoA, and α-ketoglutarate. These metabolites are actually the key branching points for carbohydrate, lipid and protein syntheses.

3.2 The yeast *Saccharomyces cerevisiae* under aerobic/anaerobic conditions

As a second example, we study the transition from aerobic to anaerobic conditions in the yeast *Saccharomyces cerevisiae*, using the metabolic network YeastGEM v8.1.1 composed of 2241 metabolites and 3520 reactions [6]. CISPER has identified as switch points pyruvate and acetaldahyde, corresponding to the junctions between respiration (TCA cycle) and fermentation (ethanol metabolism), in line with what we could expect.

4 Conclusion

A method - called CISPER - has been proposed to identify *in silico* switch nodes based on the analysis of a set of FBA solutions, corresponding to different conditions. CISPER gives sound results on simple case studies. The method is fast and scalable, e.g. it takes just a few seconds with a standard computer for a network of three thousand reactions. The main specificity of our approach is that the given set of conditions defines the node identification. Thus, the same metabolic network can be decomposed in different ways depending on which conditions are considered.

As a future work, CISPER will be compared with other methods dealing with the identification of key metabolites (e.g. [7]). The link between identified switch points and cellular regulations will also be investigated.

Acknowledgements

This work was supported by the Inria Project Lab *Algae in silico* and the ANR project *Dynalgue*. The author thanks the members of the *Algae in silico* project for fruitful discussions.

References

[1] Jeffrey D Orth, Ines Thiele, and Bernhard Ø Palsson. What is flux balance analysis? *Nature biotechnology*, 28(3):245, 2010.

[2] Caroline Baroukh, Rafael Muñoz-Tamayo, Jean-Philippe Steyer, and Olivier Bernard. Drum: a new framework for metabolic modeling under non-balanced growth. application to the carbon metabolism of unicellular microalgae. *PloS one*, 9(8):e104499, 2014.

[3] Abolfazl Rezvan and Changiz Eslahchi. Comparison of different approaches for identifying subnetworks in metabolic networks. *Journal of bioinformatics and computational biology*, 15(06):1750025, 2017.

[4] Laurent Heirendt, Sylvain Arreckx, Thomas Pfau, Sebastian N Mendoza, Anne Richelle, Almut Heinken, Hulda S Haraldsdottir, Jacek Wachowiak, Sarah M Keating, Vanja Vlasov, et al. Creation and analysis of biochemical constraint-based models using the COBRA toolbox v. 3.0. *Nature protocols*, page 1, 2019.

[5] F. Mairet, O. Bernard, P. Masci, T. Lacour, and A. Sciandra. Modelling neutral lipid production by the microalga *Isochrysis* aff. *galbana* under nitrogen limitation. *Bioresource Technology*, 102:142–149, 2011.

[6] Markus J Herrgård, Neil Swainston, Paul Dobson, Warwick B Dunn, K Yalcin Arga, Mikko Arvas, Nils Blüthgen, Simon Borger, Roeland Costenoble, Matthias Heinemann, et al. A consensus yeast metabolic network reconstruction obtained from a community approach to systems biology. *Nature biotechnology*, 26(10):1155, 2008.

[7] Julie Laniau, Clémence Frioux, Jacques Nicolas, Caroline Baroukh, Maria-Paz Cortes, Jeanne Got, Camille Trottier, Damien Eveillard, and Anne Siegel. Combining graph and flux-based structures to decipher phenotypic essential metabolites within metabolic networks. *PeerJ*, 5:e3860, 2017.

CONSENT: Scalable self-correction of long reads with multiple sequence alignment

Pierre Morisse[1], Camille Marchet[2], Antoine Limasset[2], Thierry Lecroq[1] and Arnaud Lefebvre[1]
[1] Normandie Univ, UNIROUEN, LITIS, 76000 Rouen, France
[2] Univ. Lille, CNRS, UMR 9189 - CRIStAL

Corresponding author: pierre.morisse2@univ-rouen.fr

Abstract *Third generation sequencing technologies such as Pacific Biosciences and Oxford Nanopore Technologies allow the sequencing of long reads of tens of kbs, that are expected to solve various problems, such as contig and haplotype assembly, scaffolding, and structural variant calling. However, they also reach high error rates of 10 to 30%, and thus require efficient error correction. As first long reads sequencing experiments produced reads displaying error rates higher than 15% on average, most methods relied on the complementary use of short reads data to perform correction, in a hybrid approach. However, these sequencing technologies evolve fast, and the error rate of the long reads is now capped at around 10-12%. As a result, self-correction is now frequently used as a first step of third generation sequencing data analysis projects. As of today, efficient tools allowing to perform self-correction of the long reads are available, and recent observations suggest that avoiding the use of second generation sequencing reads could bypass their inherent bias. We introduce CONSENT, a new method for the self-correction of long reads that combines different strategies from the state-of-the-art. In particular, a multiple sequence alignment strategy is combined to the use of local de Bruijn graphs. Moreover, the multiple sequence alignment benefits from an efficient segmentation strategy based on k-mers chaining, allowing to greatly reduce its time footprint. Our experiments show that CONSENT compares well to or outperforms the latest state-of-the-art self-correction methods, on real ONT datasets. In particular, they show that CONSENT is the only method able to scale to a human dataset containing ONT ultra-long reads, reaching lengths up to 340 kbp. CONSENT is freely available at https://github.com/morispi/CONSENT.*

Keywords long reads, error correction, self-correction, multiple sequence alignment

1 Introduction

Third generation sequencing technologies Pacific Biosciences (PacBio) and Oxford Nanopore Technologies (ONT) became widely used since their inception in 2011. In contrast to second generation technologies, producing reads reaching lengths of a few hundred base pairs, they allow the sequencing of much longer reads (10 kbp on average [1], and up to 1 million bp [2]). These long reads are expected to solve various problems, such as contig and haplotype assembly, scaffolding, and structural variant calling. They are however very noisy, and reach error rates of 10 to 30%, whereas second generation short reads usually display error rates of 1%. The error profiles of these long reads are also much more complex than those of the short reads, as they are mainly composed of insertions and deletions, while short reads are mostly composed of substitutions. Moreover, ONT *reads* also suffer from homopolymer bias, and contain systematic errors in these regions. As a result, error correction is often necessary, as a first step of projects dealing with long reads. As the error profiles and error rates of the long reads are much different than those of the short reads, correcting long reads requires specific algorithmic developments.

The error correction of long reads has been tackled by two main approaches. The first approach, hybrid correction, makes use of additional short reads data to perform correction. The second approach, self-correction, aims at correcting the long reads solely based on the information contained in their sequences. Hybrid correction methods rely on different techniques such as the alignment of short reads to the long reads, de Bruijn graph exploration, alignement of contigs built from the short reads to the long reads, or even Hidden Markov Models. Self-correction methods usually build around the alignment of the long reads against each other.

As first long reads sequencing experiments displayed high error rates ($> 15\%$ on average), most methods relied on the additional use of short reads data. As a result, hybrid correction used to be much more studied and much more developed. Indeed, in 2014, 5 hybrid correction tools (PBcR [3], LSC [4], ECTools [5], LoRDEC [6], Proovread [7]) and only 2 self-correction tools (PBcR-BLASR [8], PBDAGCon [9]) were available. However, third generation sequencing technologies evolve fast, and now manage to produce long reads reaching error rates

of 10-12%. Moreover, long read sequencing technologies' evolution also allows to produce higher throughputs of data, at a reduced cost, and consequently such data became more widely available. Thus, self-correction is now frequently used as a first step of data analysis projects dealing with long reads.

1.1 Related works

Due to the fast evolution of third generation sequencing technologies, and to the lower error rates they now reach, various efficient self-correction methods were recently developed. Most of them share the common first step of computing overlaps between the long reads. This can be done via a mapping approach, which only provides the positions of the similar regions of the long reads (Canu [10], MECAT [11], FLAS [12]), or via alignment, which provides the positions of the similar regions, and their actual base to base correspondence in terms of matches, mismatches, insertions, and deletions (PBDAGCon [9], PBcR [8], Daccord [13]). A directed acyclic graph (DAG) is then usually built, in order to summarize the 1V1 alignments and compute consensus, after recomputing actual alignments of mapped regions, if necessary. Other methods rely on de Bruijn graphs, either built from small windows of the alignments (Daccord), or directly from the long reads sequences with no alignment or mapping step at all (LoRMA [14]).

However, methods relying on direct alignment of the long reads are prohibitively time and memory consuming, and current implementations thus do not scale to large genomes. Methods solely relying on de Bruijn graphs and avoiding the alignment step altogether usually require deep long reads coverage, as the graphs are built for large values of k. As a result, methods relying on a mapping strategy constitute the core of the current state-of-the-art for long read self-correction.

1.2 Contribution

We present CONSENT, a new self-correction method that combines different efficient approaches from the state-of-the-art. CONSENT indeed starts by computing multiple sequence alignments between overlapping regions of the long reads, in order to compute consensus sequences. These consensus sequences are then further polished with the help of local de Bruijn graphs, in order to correct remaining errors, and reduce the final error rate. Moreover, unlike other current state-of-the-art methods, CONSENT computes actual multiple sequence alignments, using a method based on partial order graphs [15]. In addition, we introduce an efficient segmentation strategy based on k-mers chaining, which allows to reduce the time footprint of the multiple sequence alignments. This segmentation strategy thus allows to compute scalable multiple sequence alignments. In particular, it allows CONSENT to efficiently scale to ONT ultra-long reads.

Our experiments show that CONSENT compares well to the latest state-of-the-art self-correction methods, and even outperforms them on real ONT datasets. In particular, they show that CONSENT is the only method able to scale to a human dataset containing ONT ultra-long reads, reaching lengths up to 340 kbp.

2 Methods

2.1 Overview

CONSENT takes as input a FASTA file of long reads, and returns a set of corrected long reads, reporting corrected bases in uppercase, and uncorrected bases in lowercase. Like most efficient methods, CONSENT starts by computing overlaps between the long reads using a mapping approach. These overlaps are computed using an external program, and not by CONSENT itself. This way, only matched regions need to be further aligned in order to compute consensus. These matched regions are further divided into smaller windows, that are aligned independently. The alignment of these windows is performed via a multiple sequence alignment strategy based on partial order graphs. This multiple sequence alignment is realized by iteratively constructing and adding sequences to a DAG. It also benefits from an efficient heuristic, based of k-mers chaining, allowing to reduce the time footprint of computing multiple sequence alignment between noisy sequences. The DAG is then used to compute the consensus of a given window. Once a consensus has been computed, a second step makes use of a local de Bruijn graph, built from the window's sequences, in order to polish the consensus. This allows to further correct weakly supported regions, that are, regions containing weak k-mers, and thus reduce the final error rate of the consensus. Finally, the consensus is realigned to the read, and correction is performed for each window. CONSENT's workflow is summarized in Figure 1.

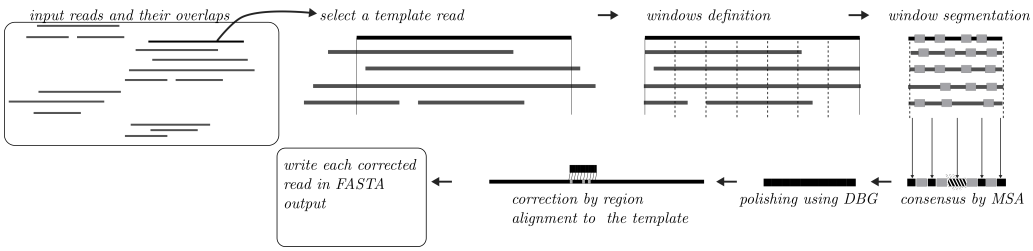


Fig. 1. Overview of CONSENT's worfklow for long read error correction.

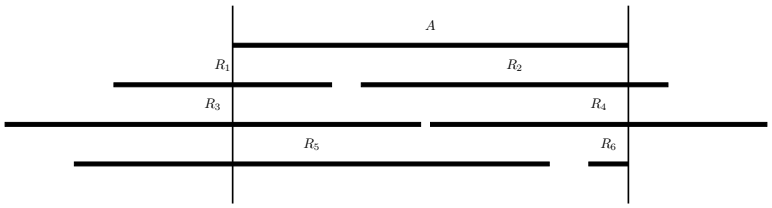


Fig. 2. An alignment pile for a template read A. The pile is delimited by vertical lines at the extremities of A. Prefixes and suffixes of reads overlapping A outside of the pile are not considered during the next steps, as the data they contain will not be useful for correcting A.

2.2 Alignment piles and windows

Before presenting the CONSENT pipeline, we recall the notions of alignment piles and windows on such piles, as proposed in Daccord, as they will be used throughout the rest of the paper.

An alignment pile represents a set of reads that overlap with a given read A. More formally, it can be defined as follows. For any given read A, we define an alignment pile for A as a set of alignment tuples (A,R,Ab,Ae,Rb,Re,S) where R is a long read id, Ab and Ae represent respectively the start and the end positions of the alignment on A, Rb and Re represent respectively the start and the end positions of the alignment on R, and S indicates whether R aligns forward (0) or reverse complement (1) to A. One can remark that, compared to Daccord, this definition is slightly altered. In particular, Daccord adds an edit script to the pile, representing the sequence of edit operations needed to transform $A[Ab..Ae]$ into $R[Rb..Re]$ if $S=0$, or into $\overline{R}[Rb..Re]$ if $S=1$ (where $\overline{R}$ represents the reverse-complement of read R). This edit script can easily be retrieved by Daccord, as it relies on DALIGNER [16] to compute actual alignments between the long reads. However, as CONSENT relies on a mapping strategy, it does not have access to such an information, and we thus chose to exclude the edit script from the definition of a pile. In its alignment pile, we call the read A the *template read*. The alignment pile of a given template read A thus contains all the necessary information needed for its correction. An example of an alignment pile is given in Figure 2.

In addition to the alignment piles principle, Daccord also underlined the interest of processing windows from these piles instead of processing them all at once. A window from an alignment pile is defined as follows. Given an alignment pile for a template read A, a window of this pile is a couple (W_b,W_e), where W_b and W_e represent respectively the start and the end positions of the window on A, and such that $0 \leq W_b \leq W_e < |A|$, *i.e.* the start and end positions of the window define a factor of the template read A. We refer to this factor as the *window's template*. Additionally, in CONSENT, we will only consider for correction windows that have the two following properties:

— $W_e - W_b + 1 = L$ (*i.e.* windows have a fixed size).
— $\forall i$, $W_b \leq i \leq W_e$, $A[i]$ is supported by at least C reads of the pile, including A (*i.e.* windows have a minimum coverage threshold).

This second property allows to ensure that CONSENT has sufficient evidence to compute reliable consensus for a window. Examples of windows considered and not considered are shown in Figure 3.

In the case of Daccord, this window strategy allows to build local de Bruijn graphs with small values of k, thus overcoming the high error rates of the long reads, causing issues with large values of k. More generally, processing windows instead of whole alignment piles allows to divide the correction problem into smaller subproblems that can be solved faster. Specifically, in our case, as we seek to correct long reads by computing multiple alignment of sequences, working with windows allows to save both time and memory, since the sequences that need to be aligned are significantly shorter.

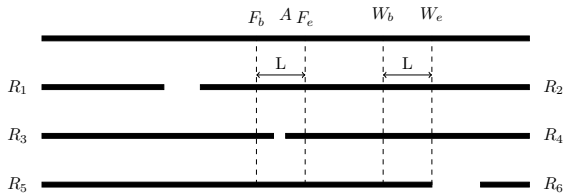


Fig. 3. When fixing the length to L and the minimum coverage threshold to 4, the window (W_b,W_e) will be processed by CONSENT. With these same parameters, the window (F_b,F_e) will not be processed by CONSENT, as $A[i]$ is not supported by at least 4 reads $\forall F_b \leq i \leq F_e$.

2.3 Overlapping

To avoid prohibitive computation time and memory consuming full alignments, CONSENT starts by overlapping the long reads using a mapping approach. By default, this step is performed with the help of Minimap2 [17]. However, error correction by CONSENT is not dependent on Minimap2, and the user can easily use any method for computing the overlaps between the long reads, as long as the overlaps file is provided to CONSENT in PAF format. We only included Minimap2 as the default overlapper for CONSENT as it offers good performances, both in terms of runtime and memory consumption, and is thus able to scale to large organisms on reasonable setups.

2.4 Alignment piles and windows computation

The alignment piles are computed by parsing the PAF file provided by the overlapper during the previous step. Each line of the file indeed contains all the necessary information to define a tuple from an alignment pile: the ids of the two mapped long reads, the start and the end positions on the two reads, as well as the strand of the second read relatively to the first.

Given an alignment pile for a read A, we can then compute its set of windows. To this aim, we use an array of the length of A, allowing to count how many times each nucleotide of A is supported. The array is initialized with 1s at each position, and for each tuple (A,R,Ab,Ae,Rb,Re,S), values at positions i such as $Ab \leq i \leq Ae$ are incremented. After all the tuples have been processed, the positions of the piles are retrieved by finding, in the array, sketches of length L of values $\geq C$, as we only consider windows of fixed length and with a minimum coverage threshold for correction. In practice, extracting overlapping windows instead of partitioning the pile into a set of non-overlapping windows has proven to be efficient. This can be explained by the fact that, due to the multiple sequence alignment performed with the windows' sequences, consensus sequence might be missing at the extremities of certain windows, as it is usually harder to exploit alignments located on sequences extremities. Such events would thus cause a lack of correction on the read, when using non-overlapping windows. Each window is then processed independently during the next steps. Moreover, the reads are loaded into memory to support random access and thus accelerate the correction process. Each base is encoded using 2 bits in order to reduce memory usage. The memory consumption is thus roughly 1/4 of the total size of the reads.

2.5 Window consensus

The processing of a window is performed in two distinct steps. First, the sequences from the window are aligned using a multiple sequence alignment strategy based on partial order graphs, in order to compute consensus. This multiple sequence alignment strategy also benefits from an efficient heuristic, based on k-mers chaining, allowing to decompose the global problem into smaller instances, thus reducing both time and memory consumption. Second, once the consensus of the window has been computed, it is further polished with the help of a local de Bruijn graph, at the scale of the window, in order to get rid of the few errors that might remain despite consensus computation.

In order to compute the consensus of a window, CONSENT uses POAv2 [15], an implementation of a multiple sequence alignment strategy based on partial order graphs. These graphs are directed acyclic graphs, and are used as data structures containing all the information of the multiple sequence alignment. This way, at each step (*i.e.* at each alignment of a new sequence), the graph contains the current multiple sequence alignment result. To add a new sequence to the multiple alignment, the sequence is aligned to the DAG, using a generalization of the Smith-Waterman algorithm.

Unlike other methods that compute 1V1 alignments between the read to be corrected and other reads mapping to it, and then build a result DAG to represent the multiple sequence alignment, this strategy allows CONSENT to directly build the result DAG, during the multiple alignment. Indeed, the DAG is first initialized

with the sequence of the window's template, and is then iteratively enriched by aligning the other sequences from the window, until it becomes the final, result graph. A matrix, representing the multiple sequence alignment, is then extracted from the graph, and the consensus is computed by performing a majority voting. In the case of a tie at a given position, the nucleotide from the window's template is chosen as the consensus base.

However, computing multiple sequence alignments on hundred of bases from dozens of sequences is computationally expensive, especially when the divergence among sequences is high. To avoid the burden of building a consensus by computing full multiple sequence alignments of long sequences, we will search for collinear regions on these sequences, in order to split the global task into several smaller instances. Several consensus will thus be built on regions delimited by anchors shared among the sequences, and the global consensus will be reconstructed from the different, smaller corrected sequences obtained. The rationale is to benefit from the knowledge that all the sequences come from the same genomic area. This way we are able to, on the one hand, compute multiple sequence alignments on shorter sequences, which greatly reduces the computational cost. On the other hand, we only use related sequences to build the consensus, and therefore exclude spurious sequences. This behavior allows a massive speedup along with a gain in the global consensus quality.

To find such collinear regions, we first select k-mers that are non repeated in their respective sequences, and shared by multiple sequences. We therefore rely on dynamic programming to compute the longest anchors chain a_1,...,a_n such that:

1. $\forall i,j$ such that $1 \leq i < j \leq n$, a_i appears before a_j in every sequence that contains a_i and a_j
2. $\forall i$, $1 \leq i < n$, there is at least T reads containing a_i and a_{i+1} (with T a solidity threshold equal to 8 by default).

We therefore compute local consensuses using substrings between successive anchors among sequences that contain them, and output the global consensus:
$consensus(prefix)+a_1+consensus(]a_1,a_2[)+a_2+\cdots+consensus(]a_{n-1},a_n[)+a_n+consensus(suffix)$.

After processing a given window, a few erroneous bases might remain on the computed consensus. This might especially happen in cases where the coverage depth of the window is relatively low. We thus propose an additional polishing feature to CONSENT as a proof of concept. This allows CONSENT to further enhance the quality of the consensus, by correcting the k-mers that are weakly supported. It is related to Daccord's local de Bruijn graph correction strategy.

A local de Bruijn graph is thus built from the window's sequences, only using small, solid, k-mers. The rationale is that small k-mers allows CONSENT to overcome the classical issues encountered due to the high error rate of the long reads, when using large k values. CONSENT searches for regions only composed of weak k-mers, flanked by sketches of n (usually, $n=3$) solid k-mers. CONSENT then attempts to find a path allowing to link a solid k-mer from the left flanking region to a solid k-mer from the right flanking region. We call these k-mers *anchors*. The graph is thus traversed, in order to find a path between the two anchors, using backtracking if necessary. If a path between two anchors is found, the region containing the weak k-mers, is replaced by the sequence dictated by the path. If none of the anchors pairs could be linked, the region is left unpolished.

To polish sketches of weak k-mers located at the left (respectively right) extremity of the consensus, highest weighted edges of the graph are followed, until the length of the followed path reaches the length of the region to polish, or no edge can be followed out of the current node.

2.6 Anchor window consensus to the read

Once the consensus of a window has been computed and polished, it needs to be reanchored on the template long read. To this aim, it is realigned to the template, using an optimized version of the Smith-Waterman algorithm. To avoid time-costly alignment, the consensus is however only locally aligned around the positions of the window it has been computed from. This way, for a window (W_b,W_e) of the alignment pile of the read A, its consensus will be aligned to $A[W_b-O..W_e+O]$, where O represents the length of the overlap between consecutive windows processed by CONSENT (usually, $O=50$, although it can be user-defined). Aligning the consensus outside of the original window's extremities allows to take into account the error profile of the long reads. Indeed, as they mainly contain insertion(s) and deletion(s) errors, it is likely that the consensus computed from a window could be longer than the window it originates from, thus spanning outside of the window's extremities. In the case that alignment positions of the consensus from the ith window overlap with alignment positions of the consensus from the $(i+1)$th window, the overlapping sequences of the two consensus are computed. The one containing the largest number of solid k-mers (where the k-mer frequencies of each

sequence are computed from the window their consensus originate from) is chosen and kept as the correction. In the case of a tie, we arbitrarily chose the sequence from consensus $i+1$ as the correction. The aligned factor of the long read is then corrected by replacing it with the aligned factor of the consensus.

3 Experimental results

3.1 Impact of the segmentation strategy

Before comparing CONSENT to any state-of-the-art self-correction tool, we first validate our segmentation strategy. To this aim, we simulated a 50x coverage of long reads from *E.coli*, with a 12% error rate, using SimLoRD [18]. We then ran the CONSENT pipeline, with, and without the segmentation strategy. Results of this experiment are given in Table 1. These results were obtained with LRCstats [19], a tool designed to measure correction accuracy on simulated long reads. These results show that, in addition to being 47x faster than the regular strategy, our segmentation strategy also allows to reach slightly lower memory consumption, and slightly higher throughput and quality.

	Without segmentation	With segmentation
Throughput	214,667,382	215,693,736
Error rate (%)	0.0757	0.0722
Runtime	5h31min	7min
Memory (MB)	750	675

Tab. 1. Comparison of the results obtained by CONSENT, with and without our segmentation stategy, as reported by LRCstats. The segmentation strategy allows a 47x speed-up, and produces slightly better results.

3.2 Comparison to the state-of-the-art

We now compare CONSENT against state-of-the-art error correction methods. We include the following tools in the benchmark: Canu, Daccord, FLAS, and MECAT. We voluntarily excluded LoRMA from the comparison, as it tends to split the reads a lot, and thus to produce reads that are usually shorter than 500 bp. We also exclude hybrid error correction tools from the benchmark, as we believe it makes more sense to only compare self-correction tools against each other. All tools were ran with default or recommended parameters. For CONSENT, we used a minimum support of 4 to define a window, a window size of 500, an overlap size of 50 between the windows, a k-mer size of 9 for the chaining and the polishing, a threshold of 4 to consider k-mers as solid. Additionally, consensus were only computed for windows having a minimum number of 2 anchors.

The different tools are compared on two real ONT datasets, one 63x coverage from *D. melanogaster*, and one 29x coverage from *H. sapiens* chromosome 1, the latter containing *ultra-long reads*, reaching lengths up to 340 kbp. Further details are given in Table 2.

We evaluate how well the corrected long reads realign to the reference genome, and report how many reads were corrected, their throughput, their N50, the proportion of corrected reads that could be aligned, the average identity of the alignments, as well as the genome coverage, that is, the percentage of bases of the reference genome to which at least a nucleotide aligned. We also evaluate how well the long reads assemble, and report the number of contigs, the number of aligned contigs, the NGA50, the NGA75 and the genome coverage of the assemblies. We aligned the long reads to the reference with Minimap2, assembled them with Miniasm, and obtained statistics by parsing the output SAM file. Results are given in Table 3 and in Table 4. Runtimes and memory consumption of the different methods are also given in Table 3. All the experiments were run on cluster node equipped with 28 2.39 GHz cores and 128 GB of RAM.

On these two datasets, Daccord failed to run, as DALIGNER could not perform alignment, reporting an error upon start, and is thus not presented in the comparison. CONSENT corrected the largest number of reads, and reached the highest alignment identity on the two datasets. Its N50 was also higher than that of all the other methods, except for Canu on the *D. melanogaster* dataset. CONSENT also reached the highest throughput, and the largest genome coverage, for the two datasets. When it comes to runtime and memory consumption, MECAT outperformed all the other methods. Moreover, it reached the highest proportion of aligned reads, on the two datasets. CONSENT was however really close, as only 0.36-0.61% less reads could be aligned. Moreover, on the *H. sapiens* dataset, CONSENT was the only tool able to scale to the ultra-long reads. Indeed, other methods reported errors when attempting to correct the original dataset. As a result, in order for those methods to be able to run, we had to remove the reads that were longer than 50kbp. There were 1,824 such reads, accounting for a total number of 135,364,312 bp. However, even after removing these

Dataset	Strain	Reference sequence	Number of reads	Average length	Error rate	Coverage	Accession
D. melanogaster	BDGP Release 6	ISO1 MT/dm6	1,327,569	6,828	14.55	63x	SRX3676783
H. sapiens	GRCh38	NC_000001.11[1]	1,075,867	6,744	17.60	29x	PRJEB23027[2]

Tab. 2. Description of the ONT long reads datasets used in our experiments. [1] Only chromosome 1 was used. [2] Only reads from chromosome 1 were used.

Dataset	Corrector	Number of reads	Throughput (Mbp)	N50 (bp)	Aligned reads (%)	Alignment identity (%)	Genome coverage (%)	Total Runtime	Total Memory (MB)
D. melanogaster	Original	1,327,569	9,064	11,853	85.52	85.43	98.47	N/A	N/A
	Canu	829,965	6,993	**12,694**	98.05	95.20	97.89	14 h 04 min	**10,295**
	daccord	-	-	-	-	-	-	-	-
	FLAS	855,275	7,866	11,742	95.65	94.99	98.09	10 h 18 min	18,820
	MECAT	849,704	7,288	11,676	**99.87**	96.52	97.34	**1 h 54 min**	13,443
	CONSENT	**1,065,621**	**8,178**	12,297	99.26	**96.72**	**98.20**	38 h	10,952
H. sapiens	Original	1,075,867	7,256	10,568	88.24	82.40	92.46	N/A	N/A
	Canu[1]	-	-	-	-	-	-	-	-
	daccord[1]	-	-	-	-	-	-	-	-
	FLAS[1]	670,708	5,695	10,198	99.06	91.00	92.37	4 h 57 min	14,957
	MECAT[1]	667,532	5,479	10,343	**99.95**	91.69	91.44	**1 h 53 min**	11,075
	CONSENT	**869,462**	**6,349**	**10,839**	99.59	**93.00**	**92.40**	8 h 30 min	**10,022**

Tab. 3. Statistics of the real long reads, before and after correction with the different methods. [1] Reads longer than 50kbp were filtered out, as ultra-long reads caused the programs to stop with an error. There were 1,824 such reads in the original datasets, accounting for a total number of 135,364,312 bp. Daccord could not be run on these two datasets, due to errors reported by DALIGNER. Canu stopped with an error on the *H. sapiens* dataset, both with and without the long reads > 50kbp.

Dataset	Corrector	Number of contigs	Aligned contigs	NGA50	NGA75	Genome coverage (%)
D. melanogaster	Original	423	408	864,011	159,590	83.1900
	Canu	410	381	**2,757,690**	**822,577**	92.1034
	Daccord	-	-	-	-	-
	FLAS	374	361	1,123,351	364,884	92.1105
	MECAT	**308**	**307**	1,425,566	478,877	89.5839
	CONSENT	455	448	1,666,202	470,720	**92.5688**
H. sapiens	Original	201	188	1,025,355	247,806	77.5700
	Canu	-	-	-	-	-
	Daccord	-	-	-	-	-
	FLAS	237	**237**	1,698,601	289,968	88.4068
	MECAT	249	247	1,672,967	424,788	88.7002
	CONSENT	**182**	177	**2,663,412**	**439,178**	**88.9587**

Tab. 4. Statistics of the assemblies obtained with the corrected long reads. As previously mentioned, Daccord results on the two datasets, and Canu results on the *H. sapiens* dataset are absent, as the tools could not be run.

ultra-long reads, Canu and Daccord still failed to perform correction, and reported errors.

The long reads corrected by CONSENT also generated satisfying assemblies, displaying the highest genome coverage on the two datasets. Moreover, on the *H. sapiens* dataset, CONSENT corrected long reads generated the assembly composed of the smallest number of contigs, and displaying the largest NGA50 and NGA75.

4 Conclusion

We presented CONSENT, a new self-correction method for long reads that combines different efficient strategies from the state-of-the-art. CONSENT starts by dividing overlapping regions of the long reads into smaller windows, in order to compute multiple sequence alignments, and consensus sequences of these windows. These multiple sequence alignments are performed using a method based on partial order graphs, allowing to perform actual multiple sequence alignment. This method is combined to an efficient k-mer chaining strategy, which allows to further divide the multiple sequence alignment into smaller instances, and thus reach greater speed. Once the consensus of a window from a matched region has been computed, it is further polished with the help of a local de Bruijn graph, in order to further reduce the final error rate, and is realigned to the read.

Our experiments show that CONSENT compares well, or even outperforms other state-of-the-art methods

in terms of quality of the results. In particular, CONSENT is the only method able to scale to a human dataset containing ONT ultra-long reads, reaching lengths up to 340 kbp. Although recent, such reads are expected to further develop, and become more accessible in the near future. Being able to deal with them will thus soon become a necessity. CONSENT could therefore be the first self-correction method able to be applied to such ultra-long reads on a greater scale.

The segmentation strategy introduced in CONSENT also shows that actual multiple sequence alignments techniques are applicable to long, noisy sequences. In addition to being useful for error correction, this could also be applied for in various other problems, such as during the consensus steps of assembly tools, for haplotyping, and for quantification problems. The literature about multiple sequence alignment is vast, but lacks application on noisy sequences. We believe that CONSENT could be a first work in that direction.

Acknowledgements

Part of this work was performed using computing resources of CRIANN (Normandy, France), project 2017020.

References

[1] Fritz J Sedlazeck, Hayan Lee, Charlotte A Darby, and Michael C Schatz. Piercing the dark matter: bioinformatics of long-range sequencing and mapping. *Nature Reviews Genetics*, page 1, 2018.

[2] Miten Jain, Sergey Koren, Karen H Miga, Josh Quick, Arthur C Rand, Thomas A Sasani, John R Tyson, Andrew D Beggs, Alexander T Dilthey, Ian T Fiddes, et al. Nanopore sequencing and assembly of a human genome with ultra-long reads. *Nature biotechnology*, 36(4):338, 2018.

[3] Sergey Koren, Michael C Schatz, Brian P Walenz, Jeffrey Martin, Jason T Howard, Ganeshkumar Ganapathy, Zhong Wang, David A Rasko, W Richard McCombie, Erich D Jarvis, and Adam M Phillippy. Hybrid error correction and de novo assembly of single-molecule sequencing reads. *Nature biotechnology*, 30(7):693–700, 2012.

[4] Kin Fai Au, Jason G. Underwood, Lawrence Lee, and Wing Hung Wong. Improving PacBio Long Read Accuracy by Short Read Alignment. *PLoS ONE*, 7(10):1–8, 2012.

[5] Hayan Lee, James Gurtowski, Shinjae Yoo, Shoshana Marcus, W. Richard McCombie, and Michael Schatz. Error correction and assembly complexity of single molecule sequencing reads. *bioRxiv*, page 006395, 2014.

[6] Leena Salmela and Eric Rivals. LoRDEC: Accurate and efficient long read error correction. *Bioinformatics*, 30:3506–3514, 2014.

[7] Thomas Hackl, Rainer Hedrich, Jörg Schultz, and Frank Förster. Proovread: Large-scale high-accuracy PacBio correction through iterative short read consensus. *Bioinformatics*, 30(21):3004–3011, 2014.

[8] Sergey Koren, Gregory P Harhay, Timothy P L Smith, James L Bono, Dayna M Harhay, Scott D Mcvey, Diana Radune, Nicholas H Bergman, and Adam M Phillippy. Reducing assembly complexity of microbial genomes with single-molecule sequencing. *Genome Biology*, 14(9):R101, sep 2013.

[9] Chen-Shan Chin, David H Alexander, Patrick Marks, Aaron A Klammer, James Drake, Cheryl Heiner, Alicia Clum, Alex Copeland, John Huddleston, Evan E Eichler, Stephen W Turner, and Jonas Korlach. Nonhybrid, finished microbial genome assemblies from long-read SMRT sequencing data. *Nature Methods*, 10:563–569, 2013.

[10] Sergey Koren, Brian P Walenz, Konstantin Berlin, Jason R Miller, Nicholas H Bergman, and Adam M Phillippy. Canu: scalable and accurate long-read assembly via adaptive k-mer weighting and repeat separation. *Genome Research*, 27:722–736, 2017.

[11] Chuan Le Xiao, Ying Chen, Shang Qian Xie, Kai Ning Chen, Yan Wang, Yue Han, Feng Luo, and Zhi Xie. MECAT: Fast mapping, error correction, and de novo assembly for single-molecule sequencing reads. *Nature Methods*, 14(11):1072–1074, 2017.

[12] Changjin Song, Ergude Bao, Fei Xie, and Dandan Song. FLAS: fast and high throughput algorithm for PacBio long read self-correction. *Bioinformatics*, btz206, 10.1093/bioinformatics/btz206.

[13] German Tischler and Eugene W Myers. Non Hybrid Long Read Consensus Using Local De Bruijn Graph Assembly. *bioRxiv, doi: https://doi.org/10.1101/106252*, 2017.

[14] Leena Salmela, Riku Walve, Eric Rivals, and Esko Ukkonen. Accurate selfcorrection of errors in long reads using de Bruijn graphs. *Bioinformatics*, 33:799–806, 2017.

[15] Christopher Lee, Catherine Grasso, and Mark F Sharlow. Multiple sequence alignment using partial order graphs. *Bioinformatics*, 18(3):452–464, 2002.

[16] Gene Myers. Efficient local alignment discovery amongst noisy long reads. In Dan Brown and Burkhard Morgenstern, editors, *Algorithms in Bioinformatics*, pages 52–67, Berlin, Heidelberg, 2014. Springer Berlin Heidelberg.

[17] Heng Li. Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics*, 34(18):3094–3100, 2018.

[18] Bianca K. Stöcker, Johannes Köster, and Sven Rahmann. SimLoRD: Simulation of Long Read Data. In *Bioinformatics*, volume 32, pages 2704–2706, 2016.

[19] Sean La, Ehsan Haghshenas, and Cedric Chauve. LRCstats, a tool for evaluating long reads correction methods. *Bioinformatics*, 33:3652–3654, 2017.

Genotyping Structural Variations using Long Reads data

Lolita LECOMPTE[1], Pierre PETERLONGO[1], Dominique LAVENIER[1] and Claire LEMAITRE[1]
Univ Rennes, CNRS, Inria, IRISA - UMR 6074, F-35000 Rennes, France

Corresponding author: lolita.lecompte@inria.fr

Abstract *Studies on structural variants (SV) are expanding rapidly. As a result, and thanks to third generation sequencing technologies, more and more SVs are discovered, especially in the human genome. At the same time, for several applications such as clinical diagnoses, it becomes important to genotype newly sequenced individuals on well defined and characterized SVs. Whereas many SV genotypers have been developed for short read data, there still have no approaches to assess whether some SVs are present or not in a new sequenced sample of long reads, from third generation sequencing technologies, such as Pacific Biosciences or Nanopore.*

In this work, we present a novel method to genotype known SVs from long read sequencing. The principle of our method is based on the generation of a set of reference sequences that represent the two alleles of each structural variant. Alignments are built from mapping the long reads to these reference sequences. They are then analyzed and filtered out to keep only informative ones, in order to quantify and estimate the presence of each allele. Currently, the genotyping of large deletions have been investigated. Tests on simulated long reads based on 1000 deletions from dbVAR show a precision of 95.8%. We also applied the method to the whole NA12878 human genome.

Keywords Structural Variations, Genotyping, Long Reads

1 Introduction

Structural variations (SV) are characterized as genomic segments of a least 50 base pairs (bp) long, that are rearranged in the genome. There are several types of SV such as deletions, insertions, duplications, inversions, translocations. With the advent of Next Generation Sequencing (NGS) and the re-sequencing of many individuals in populations, SVs have been shown to be a key component of polymorphism [1]. This kind of polymorphism have been shown involved in many biological processes such as diseases or evolution [2]. Databases referencing such variants grow as new variants are discovered, at this time dbVar, the reference database of human genomic SVs [3], contains 35,428,724 variant calls, illustrating that many SVs have already been discovered and characterized in the human population.

When studying the SVs of newly sequenced individuals, one can distinguish two distinct problems: discovery and genotyping. In the SV discovery problem, the aim is to identify all the variants that differentiate the given resequenced individuals with respect usually to a reference genome. In the SV genotyping problem, the aim is to evaluate if a given known SV (or set of SVs) is present or absent in the re-sequenced individual, and assess, if it is present, with which ploidy (heterozygous or homozygous). At first glance, the genotyping problem may seem included in the discovery problem, since present SVs should be discovered by discovery methods. However, in discovery algorithms, SV evidences are only investigated for present variants (ie. incorrect mappings) and not for absent ones. If a SV has not been called, we can not know if the caller missed it (False Negative) or if the variant is truly absent in this individual and this could be validated by a significant amount of correctly mapped reads in this region. Moreover, in the genotyping problem, knowing what we are looking for should make the problem simpler and the genotyping result probably more precise. With the fine characterization of a growing number of SVs in the human populations, genotyping newly sequenced individuals becomes very interesting and informative, in particular in medical diagnosis contexts.

In this work, we focus on this second problem: genotyping already known SVs in a newly sequenced sample. Such genotyping methods already exist for short reads data: for instance, SVtyper [4], SV^2

[5], Nebula [6], Malva [7]. Though short reads are often used to discover and genotype SVs, this is well known that their short size make them ill-adapted for predicting large SVs or SVs located in repeated regions. As a matter of fact, SVs are often located alongside repeated sequences such as mobile elements, resulting in mappability issues that make the genotyping problem harder when using short read data.

Third generation sequencing technology, such as Pacific Biosciences (PB) and Oxford Nanopore Technologies (ONT), can produce long reads data compared to NGS technologies. Long reads sequences have enabled many applications, including new SVs discoveries. Despite their high error rate, long reads are crucial in the study of SVs. Indeed, the size range of this data can reach a few kilobases (kb) to megabases, thus long reads can extend over rearranged SV sequences as well as over the repeated sequences often present at SV's breakpoint regions.

Following long reads technology's development, many SV discovery tools have emerged, such as Sniffles [8] and NanoSV [9]. Among these tools, some have a genotyping module that gives the frequency of alleles after calling SVs of the sequenced samples, nonetheless their required post-processing to evaluate if a set of SVs is present or not in the sample. To our knowledge there is currently no tool that can perform genotyping from a set of known SVs with long reads data. Thus, there is a need to develop accurate and efficient methods to genotype SVs with long reads data, especially in the context of clinical diagnoses.

The main contribution of this work is a novel method to genotype known SVs using long reads data. We also provide an implementation of this method in the tool named Biskoul. Biskoul was applied on simulated data of the human genome and on real data of the individual NA12878. High precision was achieved on both simulated and real data.

2 Materials and Methods

2.1 Methods

Pipeline We propose a method that aims at assigning a genotype for a set of already known SVs in a given individual sample sequenced with long reads data. In other words, the method assesses if each SV is present in the given individual, and if so, how many variant alleles it holds, ie. whether the individual is heterozygous or homozygous for the particular variant. The method is described and implemented here for only one type of SV, the deletions, but the principle can be easily generalized to other types of SVs. The method takes as input a variant file with deletion coordinates, a reference genome and the sample long read sequences. It outputs a variant file complemented with the individual genotype information for each input variant.

The principle of the method is based first on generating reference sequences that represent the two alleles of each SV. Then the sample long reads are aligned on the whole set of reference alleles. An important step of our method consists in selecting and counting only informative alignments to finally estimate the genotype for each known variant. The main steps are illustrated in Fig. 1.

Generating references Starting from a known variant file in vcf format and the corresponding reference genome, the first step consists in generating two sequences for each SV, corresponding to the two possible alleles. Deletions are sequences of the reference genome that may be absent in a given individual, they are characterized in the vcf file by a starting position on the reference genome and a length. The reference allele (allele 0) is therefore the sequence of the deletion with adjacent sequences at each side, and the alternative allele (allele 1) consists in the joining of the two previous adjacent sequences. Given that reads of several kb will be mapped on these references, the size of the adjacent sequences was set to 5,000 bp at each side, giving a 10 kb sequence for allele 1 and 10 plus the deletion size kb for allele 0.

Mapping Sequenced long reads are aligned on all previously generated references. We use Minimap2 [10] (version 2.16-r922), with default parameters, as it is a fast and accurate mapper, specifically designed for long erroneous reads.

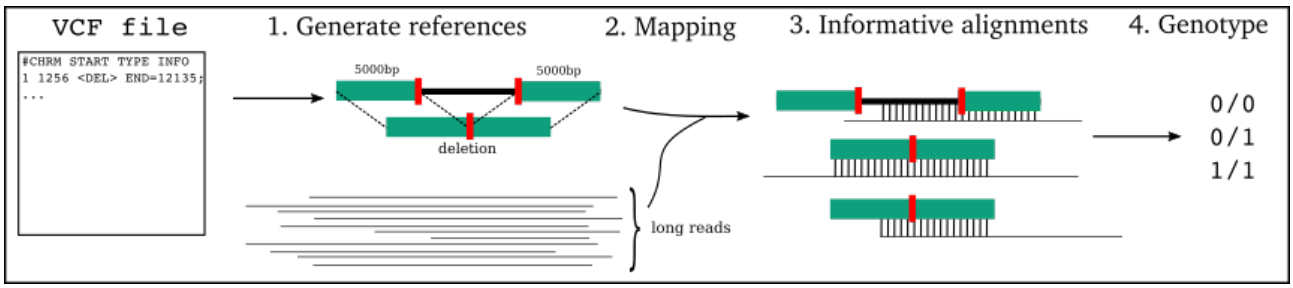


Fig. 1. Biskoul steps. 1. Two corresponding reference sequences are generated for each selected SV, one correspond to the orginal sequence and the other the sequence with the deletion. 2. Long reads sequenced data are aligned on these references using Minimap2. 3. Informative alignments are selected. 4. Genotypes are estimated.

Selecting informative alignments Minimap2 raw alignment results have to be carefully filtered out in order to remove i) uninformative alignments, that is those not discriminating between the two possible alleles, and ii) spurious false positive alignments, that are mainly due to repeated sequences.

Informative alignments for the genotyping problem are those that overlap the SV breakpoints, that is the sequence adjacencies that are specific to one or the other allele. In the case of a deletion, the reference allele contains two such breakpoints, the start and end positions of the deletion sequence; the alternative sequence, the shorter one, contains one such breakpoint at the junction of the two adjacent sequences (see the red thickmarks of Fig. 1). To be considered as overlapping a breakpoint, an alignment must cover at least d_{over} bp from each side of the breakpoint (d_{over} is set by default to 100 bp). In other words, if x and y are the distances of the breakpoint to respectively the start and end coordinates of the alignment on the reference sequence (see Fig. 2), they must satisfy the following condition in eq 1 for the alignment to be kept :

$$x \ \& \ y > d_{over} \tag{1}$$

Concerning the filtering out of spurious false positive alignments, Minimap2 alignments are first filtered based on the quality score. To focus on uniquely mapped reads, the quality score of the alignments must be greater than 10. This is not sufficient to filter out alignments due to repetitive sequences, since mapping is performed on a small subset of the reference genome and these alignments may appear as uniquely mapped on this subset.

As Minimap2 is a sensitive local aligner, many of the spurious alignments only cover subsequences of both the reference and the read sequences. To maximize the probability that the aligned read really originate from the reference locus, we therefore require that the two sequences are aligned in a semi-global manner, where each alignment extremity must correspond to an extremity of at least one of the two aligned sequences. This criteria gathers four types of situatinn, namely the read is included in the reference, or *vice-versa*, or the read left end aligns on the reference right end, or *vice-versa*. Indeed this criteria is not strictly applied and a distance of d_{end} of the alignment to an extremity is tolerated (d_{end} is set by default to 100 bp). More formally, if a and b (resp. c and d) are the distances of the alignment to the, respectively, left and right extremities of the reference sequence (resp. read sequence), then the alignment must fulfill the following condition in Eq. 2 to be kept:

$$(a < d_{end} \parallel c < d_{end}) \ \& \ (b < d_{end} \parallel d < d_{end}) \tag{2}$$

Estimating genotypes For each variant, the genotype is estimated based on the ratio of amounts of reads informatively aligned to each reference allele. Each variant has two references of different sizes, so even if both alleles are covered with the same read depth, there would be fewer reads that align on the shortest reference. To prevent a bias towards the larger allele, reported read counts for the larger

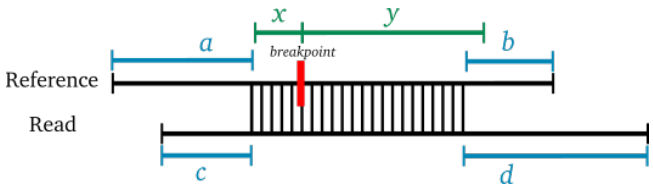


Fig. 2. Definition of the different distances of the alignment with respect to the breakpoint (x and y) and to the sequence extremities (a, b, c and d) used to select informative alignments.

alleles are normalized according to the reference sequence length ratio, assuming that read count is proportional to the sequence length.

Finally, a genotype is estimated if the variant presence or absence is supported by at least *min_cov* different reads after normalization (sum of read counts for each variant). The allele frequency is defined as the proportion of reads supporting the reference allele 0. A SV is called heterozygous, 0/1, if the allele frequency is within [0.2; 0.8]. Otherwise, if the frequency is > 0.8, the SV is called homozygous reference, 0/0, and if the frequency is < 0.2, the SV is called homozygous alternative, 1/1.

Implementation and availability We provide an implementation of this method named Biskoul. Biskoul is written in Python 3, it requires as input a set of deletion (vcf format), a genome (fasta format), a reads' file (fastq or fasta format). Also the genotyping program can be runned independently from the whole pipeline, then the user must provide a vcf file and a paf file. Biskoul is available at `https://data-access.cesgo.org/index.php/s/JhD0TNgJocVew0E`, under GNU Affero GPL licence.

2.2 Evaluation

Long reads simulated datasets Biskoul was assessed on simulated datasets of the human genome GRCh37 based on real characterized deletions for the human genome. From the dbVar database [3], we selected 1000 existing deletions on the chromosome 1, which are separated by at least 10,000 bp. The size of the deletions varies from 50 bp to 10 kbp. In this experiment, deletions were distributed into the three different genotypes: 333 deletions are considered as 0/0 genotype, 334 deletions as 0/1 genotype and the 333 remaining deletions as 1/1 genotype. We consider the homozygous 1/1 genotype, as the genotype where the deletion is present in both alleles. So, deletions were simulated on two different reference sequences, corresponding to the two haplotypes of the human genome. 1/1 genotype deletions were simulated on both reference sequences, while deletions of 0/1 genotype were randomly simulated on one of the reference sequences. Then we simulate PB long reads using SimLoRD[11] (version v1.0.2) with 16% error rate (`-pi 0.11 -pd 0.04 -ps 0.01 --max-passes 1`), at 20x depth of coverage.

Real data Biskoul was assessed on a human genome real dataset. As sequenced reads for the individual NA12878, we used ONT MinION data rel 5, from the ONT whole genome sequencing nanopore consortium data [12] (European Nucleotide Archive : PRJEB23027). ONT data were called with Guppy 0.3 (`https://s3.amazonaws.com/nanopore-human-wgs/rel5-guppy-0.3.0-chunk10k.fastq.gz`). This sequencing dataset contains 15,891,898 reads, totaling 123 Gbp, which correspond to a 39x depth of coverage.

As the set of deletions to genotype, we use the call set of variants provided by the Genome in a Bottle consortium (GiAB), for the NA12878 individual from PB data [13], (`ftp://ftp-trace.ncbi.nlm.nih.gov/giab/ftp/data/NA12878/NA12878_PacBio_MtSinai/NA12878.sorted.vcf.gz`). From this input set, only deletions sizes greater than 50 bp were used. In a second experiment we selected high confidence deletions calls from the GiAB call set.

Evaluation In order to evaluate our method, for simulated data, we compute contingency table, giving us a clear view of the number of correctly predicted genotypes as well as the number of incorrectly pre-

dicted genotypes, for each category. Also, we can assess the number of corrected predicted genotypes over all predicted genotypes, which gives the precision of the method, as shown in equation 3.

$$Precision = \frac{\text{number of correctly predicted genotypes}}{\text{number of predicted genotypes}} \quad (3)$$

3 Results

3.1 Simulated data

Biskoul was applied on PB simulated long reads for the human chromosome 1, with 1000 real characterized deletions found in dbVar, ranging from 50 to 10,000 bp. Deletions are equally distributed among the 3 different genotypes. On simulated data, Biskoul achieves 95.846% precision, it correctly predicts 942 over 987 predicted deletions among the 1,000 assessed deletions, while 13 are not estimated at all. Results are described described in Tab. 1. This simulation was repeated 10 times, and gives similar results regarding the number of correctly predicted deletions and precision.

		Prediction			
		0/0	0/1	1/1	./.
Truth	0/0	330	3	0	0
	0/1	19	305	6	4
	1/1	3	10	311	9

Precision : 95.846%

Tab. 1. Contingency table on simulated data

As we can observe in Tab. 1, the majority of false positive genotypes result from an over-mapping on reference allele 0, rather than on the alternative allele 1 (bottom left corner of the table). Indeed 19 deletions were called as 0/0 whereas they are 1/1, 10 were called as 0/1 instead of 1/1, finally 3 called as 0/0, while they are 1/1. Thus, there is a clear mapping bias towards the longest reference sequence (allele 0 contains the deletion sequence).

Interestingly, among these 32 false positives, we noticed that nearly half of them have a size less than 100 bp. This suggests that the precision of the method may depend on the deletion size. As a matter of fact, the precision is of 85.4 % for deletions smaller than 100 bp versus 97.9 % for deletions greater than 500 bp.

The remaining false positive deletions of size $\geq$ 100 bp, were manually investigated, and most of them occur in regions with a high density of mobile elements.

Comparison with SV discovery approaches One can wonder if these simulated deletions could be easily detected and genotyped by a long read SV discovery tool. We applied here the best to date such tool, Sniffles [8,14] to the chromosome 1 simulated read dataset. As expected, none of the 333 simulated deletions with 0/0 genotypes were assigned a genotype in the Sniffles output call set, since a discovery tool naturally only reports present variants. Surprisingly, among the 667 deletions simulated with either a 0/1 or 1/1 genotype, only 406 were discovered by Sniffles, which gives a recall of only 60.9 %, and with mainly the heterozygous genotypes missing (74% of 0/1 deletions were missed, versus 6 % for the homozygous ones). Interestingly, Sniffles also mis-predicts the genotype of the discovered deletions, assigning most of the 1/1 discovered deletions (n= 254, 81%) as heterozygous. This highlights the fact that Sniffles, a SV discovery tool, is much less precise for the genotyping task than a dedicated genotyping tool.

3.2 Real data

Biskoul was also applied on real ONT data for the whole human genome of NA12878 individual. We try to genotype deletions called by Genome in a Bottle (GiAB). This set of SV, refers as GiAB Mt Sinai VCF, was obtained from PB data using three different SV detection approaches, including PBHoney [15] and SMRT-SV [16]. These approaches also estimated genotypes for present variants

(heterozygous 0/1 and alternative homozygous 1/1 only), thus we can compare Biskoul prediction results with the genotype calls predicted by different methods, for this individual.

Full initial GiAB call set The input set of deletions contains 15,616 deletions, with a significant imbalance towards the heterozygous genotype with 14,185 (90%) predicted with a 0/1 genotype (deletion is present in one allele only) and 1,431 with a 1/1 genotype (deletion is present in both alleles). As expected, as a discovery result, this call set does not contain any 0/0 SV. As a result, Biskoul has assigned a genotype to 9,388 deletions, 4,388 of which were identical to the GiAB ones. As we can observe in Tab. 2, most of the differently predicted genotypes are genotyped as heterozygous in GiAB. Surprisingly, our method did not assign any genotype to 6,228 deletions, again mostly 0/1 genotyped deletions in GiAB, as indicated by the last column. This means that too few reads could be mapped to one or the other allele reference. This could be due to redundancy within the deletion call set (several closely located deletions), resulting in very similar reference sequences between variants preventing the mapper to map reads uniquely. As a matter of fact, almost half of the deletions are less than 1,000 bp apart from the preceding one.

We also predicted 3,325 deletions as 0/0, in other words, absent in NA12878. This might suggest potential False Positive calls of the GiAB call set. Finally, we notice that an important number of differently genotyped deletions, 1,607, were predicted as 1/1, whereas they were called as 0/1 in GiAB. This is in contradiction with results obtained on simulated datasets, where Biskoul errors tend to over-estimate the reference allele. This again may suggest that some deletions of the GiAB call set are mis-genotyped.

		Prediction			
		0/0	0/1	1/1	./.
GiAB call	0/0	0	0	0	0
	0/1	3,317	3,210	1,607	6,051
	1/1	8	68	1,178	177

Tab. 2. Contingency table comparing deletion genotypes of NA12878 between the GiAB full initial call set (n= 15,616) and Biskoul predictions with ONT data.

Higher confidence call set We filtered the GiAB Mt Sinai VCF in order to focus on deletions with a higher confidence call. The initial call set is the union of the deletions calls obtained with seven different discovery pipelines. Here, as a higher confidence call set, we selected the intersection set, keeping only the deletions that were detected by all seven different pipelines, which corresponds to the 'NS=1111111' flag in the INFO column. This new set of deletions contains 1,685 deletions, of which 922 are 0/1 genotypes and 763 are 1/1 genotypes. Compared to the previous experiment, we note that the ratio is more balanced between heterozygous and homozygous genotypes in this set of filtered deletions.

Biskoul was run on these selected deletions. Only one deletion could not be assigned a genotype by the method. For the 1,684 predicted deletions, 1,514 were genotyped exactly as in GiAB. This results in a much higher overlap, 89.9 %, than with the full call set. As we can observe in Tab. 3, the majority of differently predicted genotypes are 1/1 whereas they are 0/1 in GiAB call set. Again, these results are in contradiction with the evaluation of the method on simulated datasets, where the reference allele was overestimated, and therefore question the veracity of GiAB genotypes. Also previously obtained results with the SV discovery tool Sniffles, suggest a similar trend of discovery tools to mis-predict homozygous variants as heterozygous.

Performances On higher confidence GiAB call set for the human genome, with a 39 x coverage, Biskoul took 1h46m to genotype 1,687 deletions, including 1h42 for the alignment with Minimap2 parallelized on 40 cpu. Biskoul reached 6.5 Gbytes as the maximum resident set size, corresponding in fact to the memory usage of Minimap2. On the inital GiAB call set for the human genome, Biskoul took 4h02m,

		Prediction			
		0/0	0/1	1/1	./.
GiAB call	0/0	0	0	0	0
	0/1	23	780	118	1
	1/1	6	23	734	0

Tab. 3. Contingency table comparing deletion genotypes of NA12878 between the GiAB higher confidence call set (n= 1,685) and Biskoul predictions with ONT data.

to genotype 15,616 deletions, including 3h42m of mapping, and it reached a maximum memory peak of 14.7 Gbytes during mapping.

4 Discussion and Conclusion

In this work, we provide a novel SV genotyping approach for long reads data, that showed good results on both simulated and real datasets. The approach is implemented for the moment only for deletion variants in the Biskoul software. However, this proof of principle on deletion variants is a first step before generalizing the approach for all types of SVs. Insertion variants are simply the counterpart of deletions, and inversions and translocations are SVs even more balanced than insertion/deletion regarding the number of breakpoints (with exactly two breakpoints per allele). Therefore, for all these types of SVs, the method will be easily generalized, to be used in the context of clinical diagnoses or for population genomics analyses.

In the presented analyses, Biskoul ran fast within a few hours on a whole human genome dataset. Our tests show that most of the running time is dedicated to the mapping with Minimap2. Minimap2 is a fast mapper, but it spends time to compute full alignments with optimized similarity scores (ie. optimizing the locations of matches and gaps) whereas only the approximate similarity regions could be used in our approach. Thus, in order to reduce our execution time, we could use other similarity estimation strategies, such as fast alignment-free approaches [17,18].

This work also demonstrated that this is crucial to develop dedicated SV genotyping methods, as well as SV discovery methods. Firstly, because this is the only way to get evidence for the absence of SVs in a given individual. Secondly, and more surprisingly, because SV discovery tools are not as efficient and precise to genotype variants once they have been discovered, at least with long reads data as was shown here with the Sniffles experiment. Indeed, without a priori SV discovery is a much harder task than genotyping SVs with well characterized alleles, but when the aim is strictly to genotype or compare individuals on already known variants, we have shown that using as much as possible the known features of variants is much more efficient.

As a matter of fact, the efficiency of this approach depends on the quality and precision of the input variants to genotype. Although this issue is inherent to any genotyping approach, our analysis on the full GiAB call set demonstrated that our approach is probably less efficient if there is redundancy in the set of SVs to genotype. This can be frequent when the SV set is obtained from SV calling in several individuals, or with several methods as this was the case here. In these cases, this is still a difficult problem to correctly merge several call sets[19,20], and this can result in a single SV event being described by several SV entries with overlapping coordinates. This is currently not well supported by Biskoul which discards non uniquely mapped reads. The precise or rather imprecise definition of the breakpoints may also impact the genotyping performances and this remains to be assessed for this particular approach. Finally, in the perspective of applying our method for instance on the full SV catalog referenced in the dbVar database, both issues of precision and redundancy of the initial SV call set will be critical issues that may monopolize most of the efforts.

Acknowledgements

We are thankful to the Genouest bioinformatics platform, computations have been made possible thanks to the resources of the Genouest infrastructure.

References

[1] Peter A Audano, Arvis Sulovari, Tina A Graves-Lindsay, Stuart Cantsilieris, Melanie Sorensen, AnneMarie E Welch, Max L Dougherty, Bradley J Nelson, Ankeeta Shah, Susan K Dutcher, et al. Characterizing the major structural variant alleles of the human genome. *Cell*, 2019.

[2] James R Lupski. Structural variation mutagenesis of the human genome: impact on disease and evolution. *Environmental and molecular mutagenesis*, 56(5):419–436, 2015.

[3] Lon Phan, Jeffrey Hsu, Michaela Willi Le Quang Minh Tri, Tamer Mansour, Yan Kai, John Garner, John Lopez, and Ben Busby. dbvar structural variant cluster set for data analysis and variant comparison. *F1000Research*, 5, 2016.

[4] Colby Chiang, Ryan M Layer, Gregory G Faust, Michael R Lindberg, David B Rose, Erik P Garrison, Gabor T Marth, Aaron R Quinlan, and Ira M Hall. Speedseq: ultra-fast personal genome analysis and interpretation. *Nature methods*, 12(10):966, 2015.

[5] Danny Antaki, William M Brandler, and Jonathan Sebat. Sv2: accurate structural variation genotyping and de novo mutation detection from whole genomes. *Bioinformatics*, 34(10):1774–1777, 2017.

[6] Parsoa Khorsand and Fereydoun Hormozdiari. Nebula: Ultra-efficient mapping-free structural variant genotyper. *bioRxiv*, page 566620, 2019.

[7] Giulia Bernardini, Paola Bonizzoni, Luca Denti, Marco Previtali, and Alexander Schönhuth. Malva: genotyping by mapping-free allele detection of known variants. *BioRxiv*, page 575126, 2019.

[8] Fritz J Sedlazeck, Philipp Rescheneder, Moritz Smolka, Han Fang, Maria Nattestad, Arndt von Haeseler, and Michael C Schatz. Accurate detection of complex structural variations using single-molecule sequencing. *Nat Methods*, 15(6):461–468, 2018.

[9] Mircea Cretu Stancu, Markus J Van Roosmalen, Ivo Renkens, Marleen M Nieboer, Sjors Middelkamp, Joep De Ligt, Giulia Pregno, Daniela Giachino, Giorgia Mandrile, Jose Espejo Valle-Inclan, et al. Mapping and phasing of structural variation in patient genomes using nanopore sequencing. *Nature communications*, 8(1):1326, 2017.

[10] Heng Li. Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics*, 34(18):3094–3100, 2018.

[11] Bianca K Stöcker, Johannes Köster, and Sven Rahmann. Simlord: simulation of long read data. *Bioinformatics*, 32(17):2704–2706, 2016.

[12] Miten Jain, Sergey Koren, Karen H Miga, Josh Quick, Arthur C Rand, Thomas A Sasani, John R Tyson, Andrew D Beggs, Alexander T Dilthey, Ian T Fiddes, et al. Nanopore sequencing and assembly of a human genome with ultra-long reads. *Nature biotechnology*, 36(4):338, 2018.

[13] Matthew Pendleton, Robert Sebra, Andy Wing Chun Pang, Ajay Ummat, Oscar Franzen, Tobias Rausch, Adrian M Stütz, William Stedman, Thomas Anantharaman, Alex Hastie, et al. Assembly and diploid architecture of an individual human genome via single-molecule technologies. *Nature methods*, 12(8):780, 2015.

[14] Wouter De Coster, Arne De Roeck, Tim De Pooter, Svenn D'hert, Peter De Rijk, Mojca Strazisar, Kristel Sleegers, and Christine Van Broeckhoven. Structural variants identified by oxford nanopore promethion sequencing of the human genome. *BioRxiv*, page 434118, 2018.

[15] Adam C English, William J Salerno, and Jeffrey G Reid. Pbhoney: identifying genomic variants via long-read discordance and interrupted mapping. *BMC bioinformatics*, 15(1):180, 2014.

[16] Mark JP Chaisson, John Huddleston, Megan Y Dennis, Peter H Sudmant, Maika Malig, Fereydoun Hormozdiari, Francesca Antonacci, Urvashi Surti, Richard Sandstrom, Matthew Boitano, et al. Resolving the complexity of the human genome using single-molecule sequencing. *Nature*, 517(7536):608, 2015.

[17] Nicolas Maillet, Claire Lemaitre, Rayan Chikhi, Dominique Lavenier, and Pierre Peterlongo. Compareads: comparing huge metagenomic experiments. *BMC Bioinformatics*, 13(Suppl 19):S10, 2012.

[18] Camille Marchet, Lolita Lecompte, Antoine Limasset, Lucie Bittner, and Pierre Peterlongo. A resource-frugal probabilistic dictionary and applications in bioinformatics. *Discrete Applied Mathematics*, pages 1–11, April 2018.

[19] Hemang Parikh, Marghoob Mohiyuddin, Hugo YK Lam, Hariharan Iyer, Desu Chen, Mark Pratt, Gabor Bartha, Noah Spies, Wolfgang Losert, Justin M Zook, et al. svclassify: a method to establish benchmark structural variant calls. *BMC genomics*, 17(1):64, 2016.

[20] Daniel C Jeffares, Clemency Jolly, Mimoza Hoti, Doug Speed, Liam Shaw, Charalampos Rallis, Francois Balloux, Christophe Dessimoz, Jürg Bähler, and Fritz J Sedlazeck. Transient structural variations have strong effects on quantitative traits and reproductive isolation in fission yeast. *Nature communications*, 8:14061, 2017.

mCNA : a new methodology to improve high-resolution copy number variation analysis from next generation sequencing using unique molecular identifiers.

PJ Viailly[1,2,3], A. Abdel Sater[1,2], H. Dauchel[4], A. Celebi[1,2,5], C. Berard[4], N. Vergne[6], M. Viennot[1,2], E. Bohers[1,2], P. Ruminy[1,2], T. Lecroq[4], H. Tilly[1,2], P. Vera[2,4] and F. Jardin[1,2]

[1] Normandie Univ, UNIROUEN, INSERM U1245, 76000 Rouen, France
[2] Centre Henri Becquerel, 76000 Rouen, France
[3] Normandie Univ, EdN BISE 497, France
[4] Normandie Univ, UNIROUEN, LITIS EA 4108, 76000 Rouen, France
[5] Normandie Univ, UNIROUEN, Master Bioinformatique
[6] Normandie Univ, UNIROUEN, LMRS UMRS 6085, France

Corresponding author: pierre-julien.viailly@chb.unicancer.fr

Abstract *Copy number variation (CNV) of a genomic locus is one of the most important somatic aberration in the genome of tumor cells, leading to oncogene activation (gain of genomic segments) or tumor suppressor gene inactivation (deletion of genomic segments).*

Comparative genome hybridization technologies (CGH, aCGH), despite their performances, present limitations for daily clinical practice (specific platform, resolution, quantity of DNA material). For diagnostic laboratories next generation sequencing technologies (NGS) have provided the opportunities of an accurate detection of short genetic variations (single nucleotide variation, SNV). Concerning CNV, a challenging problem for the dedicated CNV detection algorithms resides in the biais of the dosage a the genomic loci, introduced by the PCR amplification step before sequencing. But, recent advances in NGS protocols offers to add unique molecular identifiers (UMIs) to each DNA molecule to be sequenced, hence providing to directly count the number of a given DNA molecule before the amplication of the library.

Here, we present mCNA (molecular Copy Number Alteration), a new methodology allowing the detection of copy number changes combining an UMI approach and a read-depth (RD) algorithm. We demonstrate the success of this approach on high and low coverage patient datasets of Diffuse Large B-Cell Lymphoma (DLBCL) cohorts, comparing to the literature. Furthermore, first results show a strong correlation with CGH detection, and an enhanced sensitivity. Finally, we demonstrate that UMI librairies used by mCNA could be useful for the detection of CNV changes not only from tissue biopsies but also from cell-free DNA samples (cfDNA). Comparison with other existing tools based on RD algorithm only is ongoing.

Keywords CNV analysis, Unique Molecular Identifiers, Next Generation Sequencing

1 Introduction

Recently, copy number variation (CNV) has gained considerable interest as a type of genomic variation that plays an important role in oncogenic pathways and disease susceptibility [1]. CNV is one of the most important somatic aberration in the genome of tumor cells. Oncogene activation and tumor suppressor gene inactivation are often attributed to copy number gain/amplification or deletion, respectively [2].

CNV analysis refers to the detection of a difference in the dosage of a genomic locus containing one or more dosage-sensitive genes (zygosity). The resolution limit of conventional cytogenetics (approximately 5 Mb) has been improved by molecular cytogenetics using comparative genomic hybridization (CGH) and more recently array comparative genomic hybridization (aCGH). These technologies make it possible to detect genomic imbalances of $< 100kb$, whereas more specialized array designs allow to increase the resolution to $< 200bp$ for specific targeted regions [3]. Despite these performances, aCGH requires the purchase of a specific platform for data acquisition and its resolution is limited to the detection of tumoral clones that differ substantially in DNA content from a reference. aCGH implies also about 100ng of DNA material limiting its use in a daily clinical practice.

Next generation sequencing technologies (NGS) have rapidly supplanted traditional Sanger sequencing as the preferred methodology for the detection of actionable single nucleotide variations (SNV) in oncology. Diagnostic laboratories are now massively equipped with Illumina/Thermofisher sequencers. Massively parallel sequencing offers many advantages including high sensitivity and specificity for SNV and CNV detection within a single platform. Nevertheless, libraries must be amplified by PCR to produce a sufficient amount of signal for next-generation sequencers. This step of amplification introduces many biases for counting reads because the number of produced reads is not anymore directly proportional to the number of initial unique targeted DNA fragments. The amplification factor of each region is unknown and depends on many parameters such as the library size, the GC content, the region length or the competition between primers overlapping the same locus.

There are three main approaches to identify CNV from NGS data: read-pair (RP), split-read (SR), and read-depth (RD) [4]. The RD approach consists in counting aligned reads overlapping a genomic region in a sliding window. These read counts (RC) are then compared between the sample of interest and a reference to compute CNV segmentation. A local decrease of sequencing depth will be associated with a loss of genomic material whereas its increase will be correlated to locus gain/amplification. This strategy looks particularly promising for the analysis of targeted sequencing experiments but removing the biases introduced during the library amplification still remain challenging.

Recent advances in NGS protocols allow to add unique molecular identifiers (UMIs) to each read. Each targeted DNA fragment is labelled by a unique random nucleotide sequence contained in sequencing primers. UMIs are especially useful for CNV detection by making each DNA molecule in a population of reads distinct. They allow to directly count the number of molecules before the amplification of the library by simply counting the number of unique UMI sequence per position of the alignment.

Here, we present mCNA (molecular Copy Number Alteration), a new methodology allowing the detection of copy number changes using a combined read-depth (RD) and UMI-based approach. We demonstrate the success of our algorithm on high and low coverage datasets of patients suffering from Diffuse Large B-Cell Lymphoma (DLBCL) and we highlight that mCNA results have a strong correlation with CGH. Finally, we demonstrate that UMI librairies analyzed with mCNA could be useful for the detection of CNV changes using formalin-fixed paraffin-embedded (FFPE) tissues and also cell-free DNA samples.
A comparison with other existing tools only based on RD algorithm is ongoing and will be discussed.

2 mCNA workflow

In this document, we present a new strategy to detect copy number changes using a combined UMI/read depth approach for targeted panels of genes. The algorithm is composed of several steps : the construction of read and UMI count matrices, the normalization of control samples to construct

a pseudo-reference, the computation of Log Ratios (LR), the segmentation and finally the statistical inference of segmented breaks using a Gaussian mixture model and conventional statistical tests.

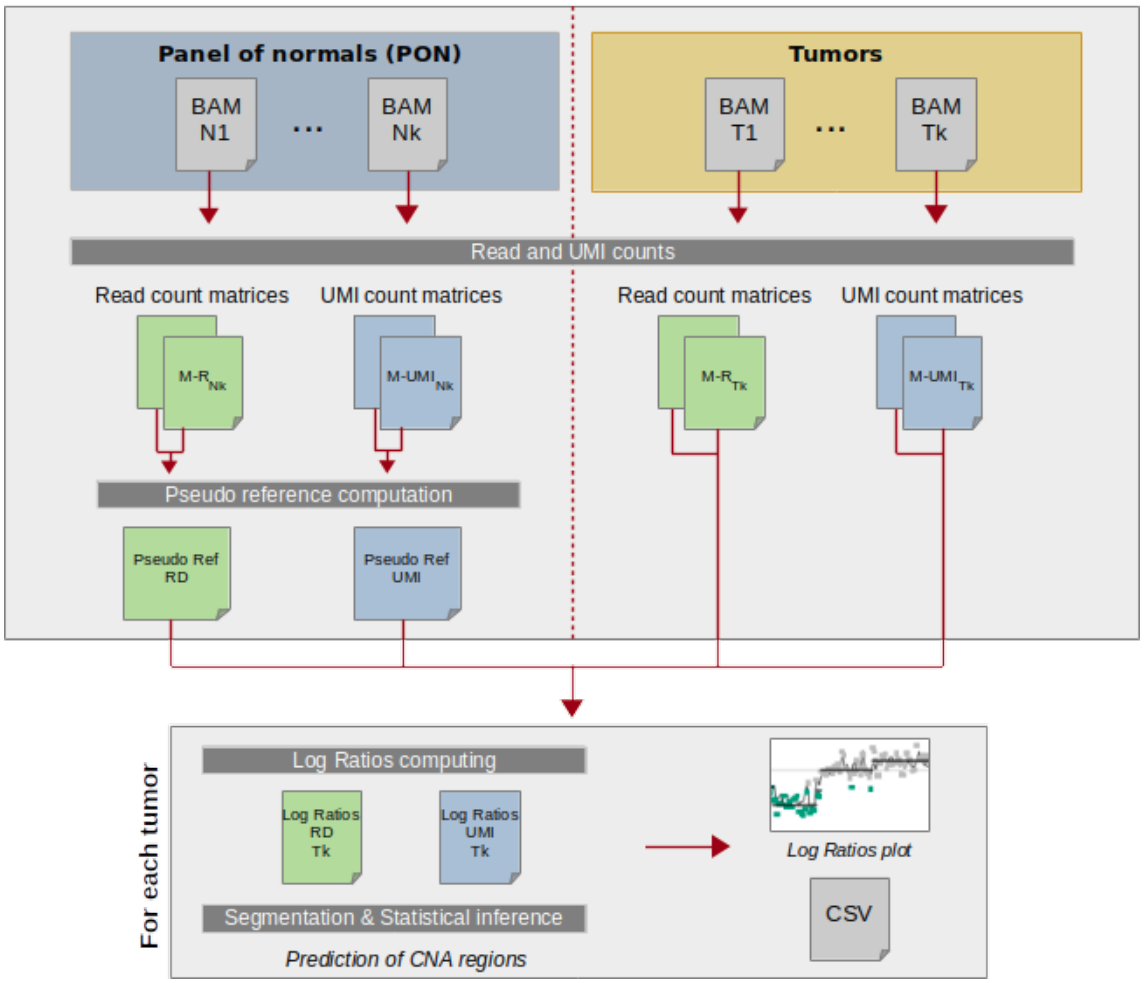


Fig. 1. Workflow of mCNA data processing. A pseudo-reference is constructed using a panel of normal samples (PON). Log ratios for both UMI counts and RD counts are computed separately. After segmentation using UMI profiles, mCNA predicts CNA regions and produces a log ratios plot.

2.1 Prerequisites

mCNA algorithm requires targeted sequencing libraries introducing one or more short aleatory sequences in reads construction. UMI sequences must be extracted from raw FASTQ files and appended to reads identifiers using UMI-tools [5]. Processed reads must then be aligned against a reference genome to produce BAM file. A BED file must be provided, giving for each targeted genomic region the chromosome name and the start/end positions of the locus.

2.2 Read/UMI counts and Pseudo-references construction

For each normal/tumoral sample and each region of the BED file, aligned reads are scanned to extract the UMI sequence from read identifiers and to compute sequencing depth per region using Pysam pileup [6]. Two matrices, M_{depth} and M_{UMI}, are thus constructed, giving the number of aligned reads and the number of unique UMI in a specific region of the alignment. In order to make the samples comparable to each other, the matrices M_{depth} and M_{UMI} are normalized, for each sample, by mean sequencing depth and mean UMI depth, respectively.

From the matrices M_{UMI} and M_{depth} of each normal sample, a geometric mean is computed to create two pseudo-references, R_{UMI} and R_{depth} respectively. These matrices will be used as reference, for each targeted region, to compute log ratios.

2.3 Log ratios and segmentation

Giving a genomic region i and a tumoral sample s, two distinct log ratios $LR^{i,s}_{UMI}$ and $LR^{i,s}_{depth}$ are computed from the matrices M_{UMI} and M_{depth} respectively, and from the two pseudo-references, R_{UMI} and R_{depth} :

$$LR^{i,s}_{UMI} = log2(\frac{M^{i,s}_{UMI}}{R_{UMI^i}}) \Big| LR^{i,s}_{depth} = log2(\frac{M^{i,s}_{depth}}{R_{depth^i}})$$

Profiles are then segmented from $LR^{i,s}_{UMI}$ using the circular binary segmentation (CBS) algorithm from DNAcopy R package [7].

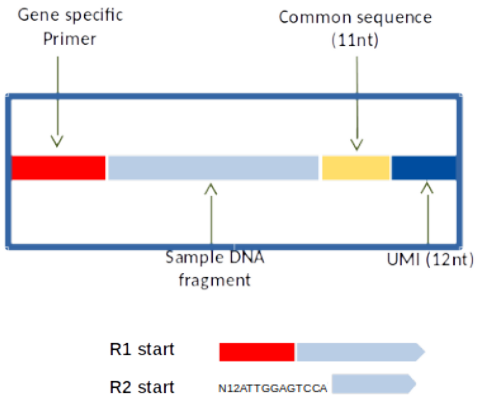
2.4 Statistical inference of segmented regions

For each CBS segment, a t-test is used to determine if there is a significant difference between the LR_{UMI} values within the segment and a theoretical normal LR value of 0. Furthermore, a clustering of LR_{UMI} using a gaussian mixture model is performed using Mclust [8] to predict outliers and to estimate the zygosity.

3 Librairies construction and patient datasets

The shown results to illustrate mCNA algorithm derived from a targeted Lymphopanel designed to identify mutations and CNV in 36 genes selected according to literature and whole exome sequencing studies of relapsed/refractory DLBCL patients [9] [10][11]. The design covers 63.700 bases using 847 gene specific primers (GSP). The libraries were prepared using QIAseq Targeted DNA Panels and include UMI of 12 random nucleotides. Samples were sequenced using a paired-end sequencing of 2x125 bp on a MiSeq (Illumina).

A.



B.

Gene	Number of GSP	Gene	Number of GSP
ARID1A	85	GNA13	14
B2M	6	ID3	6
BCL2	8	IRF4	22
BRAF	2	MEF2B	21
BTK	3	MYC	16
CARD11	16	MYD88	8
CCND3	15	NOTCH1	19
CD58	17	NOTCH2	24
CD79A	5	PIM1	14
CD79B	4	PLCG2	16
CDKN2A	27	PRDM1	33
CDKN2B	47	SOCS1	8
CIITA	66	STAT6	14
CREBBP	103	TCF3	5
CXCR4	12	TNFAIP3	31
EP300	102	TNFRSF14	16
EZH2	5	TP53	24
FOXO1	22	XPO1	11

Fig. 2. LymphoPanel : A. Structure of Qiaseq libraries. Each read of the library is composed of a Gene Specific Primer (GSP), a targeted DNA, a Common Sequence for PCR amplification and finally a random sequence of 12 nucleotides (the UMI). The UMI sequence could be extracted from R2 fastq. **B. List of targets :** List of genes of the LymphoPanel and number of associated GSP.

To compare mCNA results and other read-based published algorithms, we use a first cohort of DLBCL from the prospective, multicenter, and randomized SENIOR LYSA trial (n=150, dataset 1).

To compare mCNA and CGH performance at gene level, we use results from the prospective, multicenter, and randomized LNH-03B Lymphoma Study Association (LYSA) trial (n=20, dataset 2).

Finally, we investigate the feasibility of CNV detection on cell-free DNA samples from the prospective LymphoSeq trial at the Centre Henri Becquerel, Rouen (n=5, dataset 3).

4 Results

4.1 mCNA profile on a DLBCL biopsy (dataset 1)

Biopsies are most commonly a mixt of normal and tumoral cells. This level of contamination is often unknown and complicates the interpretation of log ratios distribution to assess the zygosity of a genomic segment. mCNA includes an unsupervised clustering using a Gaussian mixture model that assumes all the data points are issued from a mixture of a finite number of Gaussian distributions with unknown parameters. Each value of LR_{UMI}, for a targeted region, is attributed to a Gaussian, sometimes allowing to determine the zygosity of a segment and the percent of tumoral cells.

All the profiles, obtained with mCNA for all the dataset 1, were in agreement with copy number anomalies described in literature for DLBCL. For exemple, the profile on the Fig 3 typically indicates a series of well known deletions (deletions of *PRDM1*, *TNSFR14*, *CDKN2A/B* and *CREBBP*))

To deepen the performance of mNCA, comparison with other algorithms is still in progress.

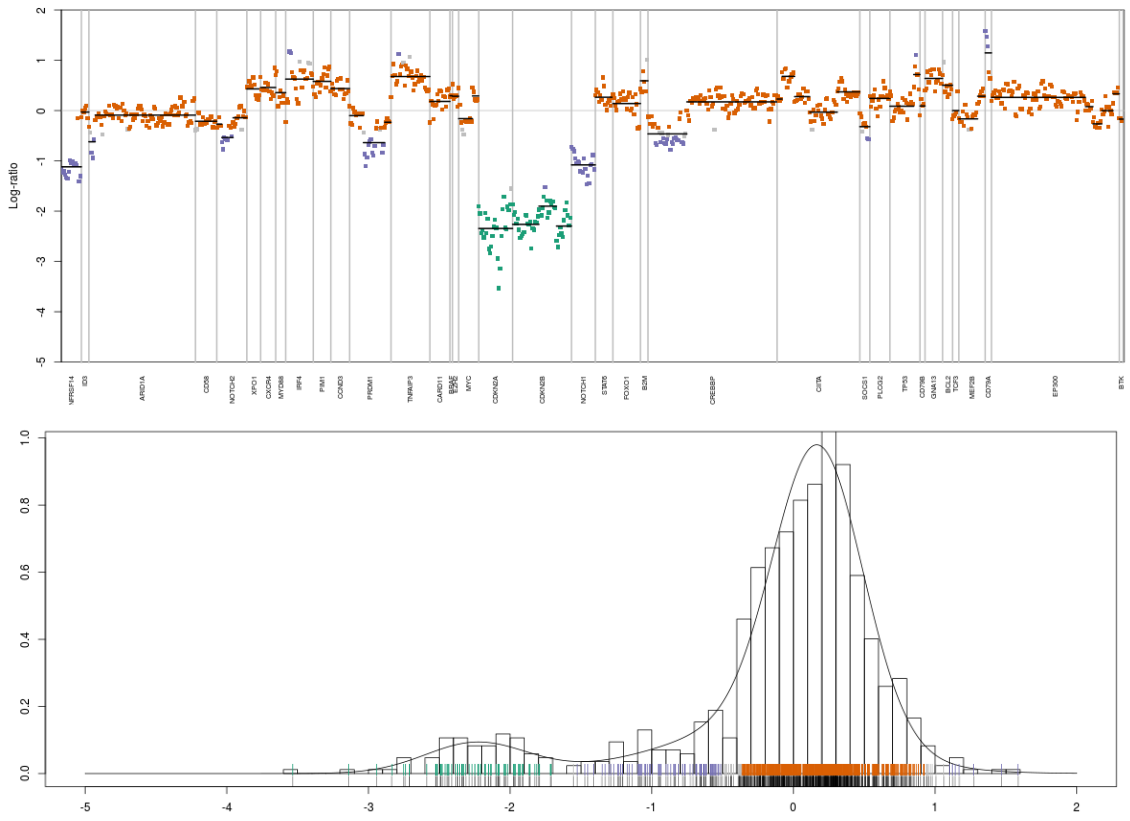


Fig. 3. Example of a mCNA profile obtained for a patient of the dataset 1. The upper part of the figure shows the results obtained using the LymphoPanel and mCNA algorithm on a biopsy of DLBCL at the time of diagnosis. Points represent LR_{UMI} throughout the targeted panel. Segments reflect CBS segmentation on LR_{UMI} values. The bottom part of the graph displays the LR_{UMI} distribution. Points were colored according to an unsupervised clustering using a Gaussian mixture model. On this profile, three gaussians were predicted corresponding to homozygous deletions (green), heterozygous deletions (purple) and normal segments (orange). Colors obtained from the clustering are reported to mCNA profile.

4.2 Comparison between mCNA and CGH (dataset 2)

In order to validate mCNA approach, we first compared log ratios obtained from CGH and NGS (Fig. 4). A strong correlation between both technologies is observed for $LR_{CGH} \neq 0$. First results seem to show that this variability is due to PCR bias (GC content) and to a lack of sensibility due to CGH's design. Preliminary results seem to show a significant correlation at gene level between mCNA and CGH results in terms of sensitivity and specificity after segmentation.

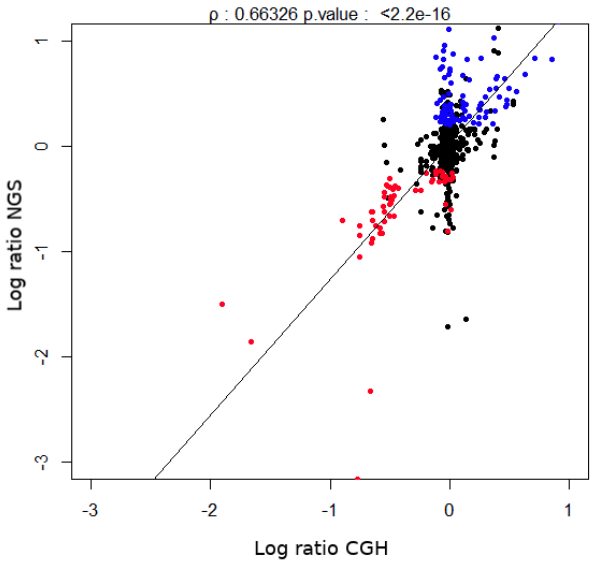


Fig. 4. Correlation between NGS and CGH : log ratios were computed from CGH signal and NGS depth. A Pearson correlation coefficiant was estimated. Points were colored according to the t-test significance and the CNV type (deletion/amplification).

Interestingly, mCNA detects short events that were missed by the CGH and by conventional variant callers ((Fig. 5). The approach seems to be able to detect deletions on the scale of only a few amplicons.

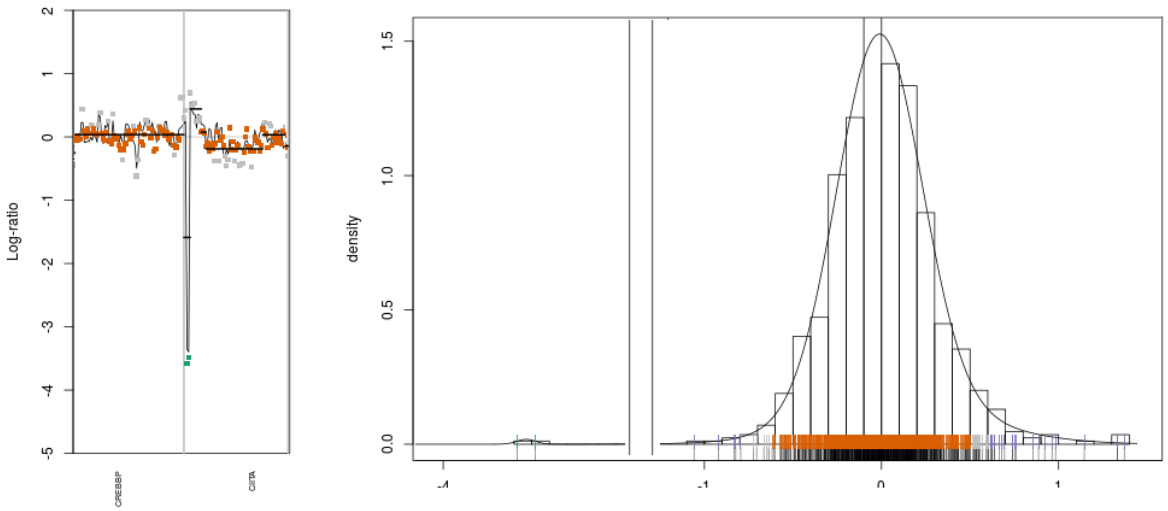


Fig. 5. Example of short deletion : The measurement of the noise is represented on this density curve by a mixture of two Gaussian distributions (orange, grey). mCNA detects an homozygous deletion of a region of $\approx 220pb$ at the beginning of the gene *CIITA* (green).

The comparison between mCNA and CGH data is still in progress.

4.3 Example of profile on cell-free DNA (dataset 3)

Cell free DNA (cfDNA) are degraded DNA fragments released to the blood plasma. Elevated levels of cfDNA are observed in cancer, especially in advanced disease. cfDNA has been shown to be a useful biomarker in cancer such as lymphoma [9].

Describing CNV in cfDNA is challenging because extracted DNA fragments are often shorter and degraded. The use of UMI instead of sequencing depth to compute log ratios looks promising because

it allows the quantification of unique DNA fragment regardless of the size of the aligned reads. We demonstrate that UMI counts and the use of mCNA algorithm can allow the detection of deletions in cfDNA samples (Fig.6).

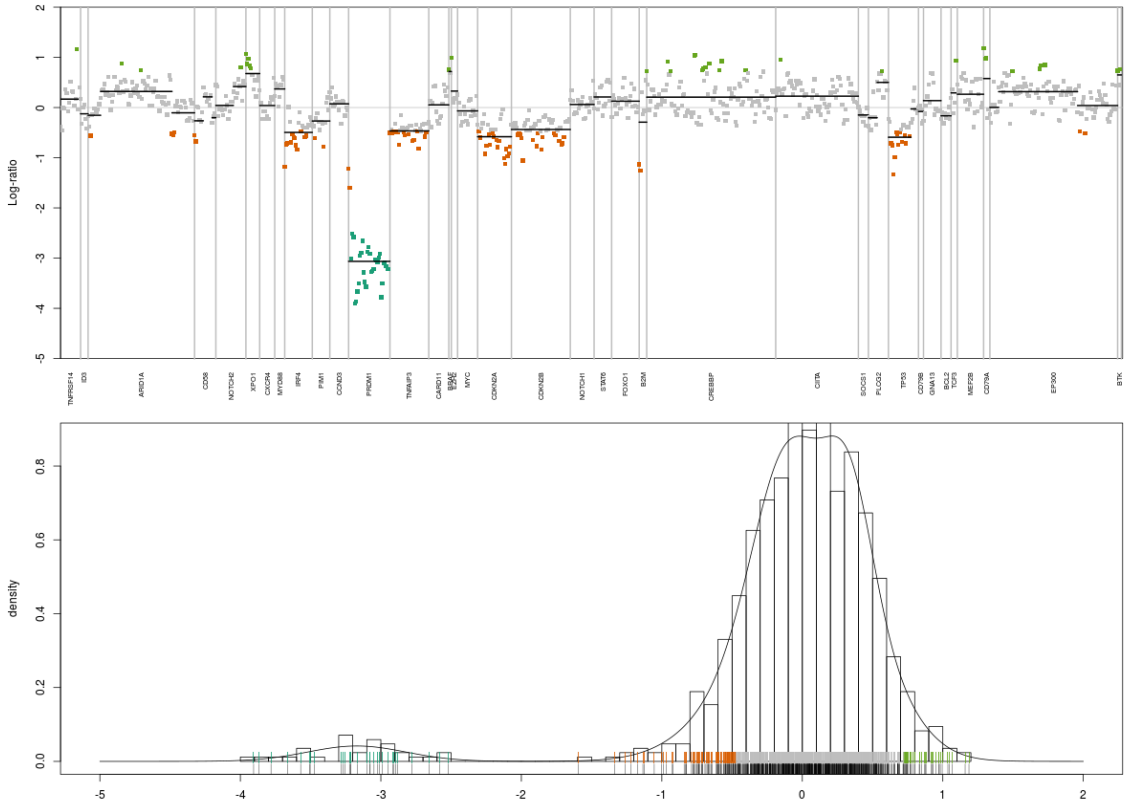


Fig. 6. Example of cfDNA profile.

Acknowledgements

The authors thank the Lymphoma Study Association and the Centre Henri Becquerel (Rouen) to provide NGS and CGH datasets.

References

[1] Fabrice Jardin, Jean-Philippe Jais, Thierry-Jo Molina, Françoise Parmentier, Jean-Michel Picquenot, Philippe Ruminy, Hervé Tilly, Christian Bastard, Gilles-André Salles, Pierre Feugier, Catherine Thieblemont, Christian Gisselbrecht, Aurelien de Reynies, Bertrand Coiffier, Corinne Haioun, and Karen Leroy. Diffuse large B-cell lymphomas with CDKN2a deletion have a distinct gene expression signature and a poor prognosis under R-CHOP treatment: a GELA study. *Blood*, 116(7):1092–1104, August 2010.

[2] Adam Shlien and David Malkin. Copy number variations and cancer. *Genome Medicine*, 1(6):62, June 2009.

[3] A Theisen. Microarray-based Comparative Genomic Hybridization (aCGH) | Learn Science at Scitable.

[4] Fatima Zare, Michelle Dow, Nicholas Monteleone, Abdelrahman Hosny, and Sheida Nabavi. An evaluation of copy number variation detection tools for cancer using whole exome sequencing data. 18(1):286.

[5] Tom Smith, Andreas Heger, and Ian Sudbery. UMI-tools: modeling sequencing errors in Unique Molecular Identifiers to improve quantification accuracy. *Genome Research*, 27(3):491–499, 2017.

[6] Heng Li, Bob Handsaker, Alec Wysoker, Tim Fennell, Jue Ruan, Nils Homer, Gabor Marth, Goncalo Abecasis, Richard Durbin, and 1000 Genome Project Data Processing Subgroup. The Sequence Alignment/Map format and SAMtools. *Bioinformatics (Oxford, England)*, 25(16):2078–2079, August 2009.

[7] VE Seshan and A Olshen. DNAcopy. *R package version 1.56.0*.

[8] Chris Fraley, Adrian E. Raftery, Luca Scrucca, Thomas Brendan Murphy, and Michael Fop. mclust: Gaussian Mixture Modelling for Model-Based Clustering, Classification, and Density Estimation, March 2019.

[9] Elodie Bohers, Pierre-Julien Viailly, Stéphanie Becker, Vinciane Marchand, Philippe Ruminy, Catherine Maingonnat, Philippe Bertrand, Pascaline Etancelin, Jean-Michel Picquenot, Vincent Camus, Anne-Lise Menard, Emilie Lemasle, Nathalie Contentin, Stéphane Leprêtre, Pascal Lenain, Aspasia Stamatoullas, Hélène Lanic, Julie Libraire, Sandrine Vaudaux, Louis-Ferdinand Pepin, Pierre Vera, Hervé Tilly, and Fabrice Jardin. Non-invasive monitoring of diffuse large B-cell lymphoma by cell-free DNA high-throughput targeted sequencing: analysis of a prospective cohort. *Blood Cancer Journal*, 8(8):74, 2018.

[10] Sydney Dubois, Pierre-Julien Viailly, Elodie Bohers, Philippe Bertrand, Philippe Ruminy, Vinciane Marchand, Catherine Maingonnat, Sylvain Mareschal, Jean-Michel Picquenot, Dominique Penther, Jean-Philippe Jais, Bruno Tesson, Pauline Peyrouze, Martin Figeac, Fabienne Desmots, Thierry Fest, Corinne Haioun, Thierry Lamy, Christiane Copie-Bergman, Bettina Fabiani, Richard Delarue, Frédéric Peyrade, Marc André, Nicolas Ketterer, Karen Leroy, Gilles Salles, Thierry J. Molina, Hervé Tilly, and Fabrice Jardin. Biological and Clinical Relevance of Associated Genomic Alterations in MYD88 L265p and non-L265p-Mutated Diffuse Large B-Cell Lymphoma: Analysis of 361 Cases. *Clinical Cancer Research: An Official Journal of the American Association for Cancer Research*, 23(9):2232–2244, May 2017.

[11] Sydney Dubois, Pierre-Julien Viailly, Sylvain Mareschal, Elodie Bohers, Philippe Bertrand, Philippe Ruminy, Catherine Maingonnat, Jean-Philippe Jais, Pauline Peyrouze, Martin Figeac, Thierry J. Molina, Fabienne Desmots, Thierry Fest, Corinne Haioun, Thierry Lamy, Christiane Copie-Bergman, Josette Brière, Tony Petrella, Danielle Canioni, Bettina Fabiani, Bertrand Coiffier, Richard Delarue, Frédéric Peyrade, André Bosly, Marc André, Nicolas Ketterer, Gilles Salles, Hervé Tilly, Karen Leroy, and Fabrice Jardin. Next-Generation Sequencing in Diffuse Large B-Cell Lymphoma Highlights Molecular Divergence and Therapeutic Opportunities: a LYSA Study. *Clinical Cancer Research: An Official Journal of the American Association for Cancer Research*, 22(12):2919–2928, 2016.

Novel insight on molecular dynamics trajectories : local equilibrium viewed by kappa-segmentation

Sharad Goulam[1], Luba Tchertanov[1] and Alain Trouvé[1]
[1] CMLA, ENS Paris-Saclay, 61 Avenue du Président Wilson, 94235, Cachan, France

Corresponding author: sharad.goulam@gmail.com

Abstract *Molecular dynamics (MD) simulations can produce nowadays huge amount of data using high-throughput CPU/GPU clusters. However, the systematic and routine use of MD simulations for a study of the large molecules of real biological systems is still considerably impeded by a lack of adequate modelling. This leads to a limited understanding of the produced highly complex signals that emerge at the level of the relevant subsystems for various time scales. We will present here an ongoing work towards the dynamics modelling and detection of local equilibrium for relevant subsystems compatible with the usual practice of MD and aiming at avoiding the detection of spurious artefactual local equilibrated states. Such well characterized local equilibrium would be basic descriptive atoms extracted from various MD trajectories.*

Keywords Molecular dynamics simulation, kinetic-based clustering, local equilibrium, segmentation, macromolecules

1 Introduction

Due to scientific and technological advancement of the past decades, the bioinformatics community is now able to generate a huge amount of data encoding molecular dynamics (MD), i.e. MD trajectories of large molecular systems on a long time scale ($\sim \mu s$ level). However, if we often analyse the generated conformational space in terms of meta-stable states, or at least, as conformational wells that correspond to local minima of the energy of a biomolecule, the existing methods and tools employed by researchers remain deficient [1]. Indeed, classical approaches such as spectral clustering or Jarvis-Patrick clustering ignore the dynamics of the problem, and do not allow to quantify, and thusly to compare the depth of the identified wells and their content. On the other hand, the widely used RMSD and RMSF methods are likely enable to consider the stable or equilibrated states, which are in reality transient segments of MD trajectory. Finally, as we do not have access to the energy of the biological macromolecule itself, but only of the global system including its environment, any method requiring a minimization of the energy of the system is not appropriate.

Here we present a new method, the *κ-segmentation*, conceived for clustering of MD-trajectory on the segments to fill the drawbacks of the previously quoted approaches. The algorithm of the *κ-segmentation* method is based on the original criteria (metrics) providing a quantification of wells depth, allowing the user to compare a well content and sort them, but also to reject any false artefacts.

2 The lap number κ: a new tool to investigate well depth

To start with, we consider the projected trajectory on the d first PCA coordinates so that we get a trajectory $(X(t))_{t\in[0,T]}$ in $\mathbb{R}^d$. Our core idea is to derive a kinetically based segmentation algorithm through the detection and analysis of regions of the configuration space into which the process spend an unexpected large time before exit.

We first the assume that trajectory as a continuous stochastic process solution of Stochastic Differential Equation (SDE) $dX_t = b(X_i) + \sigma(X_i)dB_t$ driven by a d-dimensional brownian motion. A natural idea is to compute for any time segment $[s,T] \subset [0,T]$ the radius of the smallest ball centered at X_s containing the path $u \to X_u$ for $s \leq u \leq t$:

$$R_{\max}(s,t) \doteq \sup\{\ \|X_u – X_s\|\ :\ s \leq u \leq t\ \}$$

The rational is that, in the case of a pure diffusion process, i.e $b \equiv 0$ and σ is constant) $R^2_{\max}(s,t)$ is expected to be of the order $D(t-s)$ (where D is the $D = d\sigma^2$ is the diffusion factor. It is natural to introduce the *dimensionless* quantity κ, called hereafter the lap number:

$$\kappa(s,t) \doteq \frac{D(t-s)}{R^2_{\max}(s,t)}\,. \tag{1}$$

We can expect that under the null hypothesis H_0 ($b \equiv 0$, pure diffusion), $t \to \kappa(s,t)$ stay of the order 1 (in fact, this quantity may converge to zero very slowly as $1/\log(\log(t))$ for large t due to the law of iterated logarithm). On the opposite, a large lap number may lead to the rejection of H_0 and trigger the detection of potential local equilibrium for the dynamic.

We illustrate this phenomenon in Fig. 2, comparing the evolution of $R_{\max}$ computed for two trajectories: the first one corresponding to the solution of a Langevin SDE where the drift term is the gradient of a three wells energy landscape (see [2]). Both trajectories are calibrated to have a diffusion coefficient compatible with observed ones on real MD trajectories (see the VKOR case below) and sample every 4 ps on a $T = 100$ ns simulation time. Despite having the same diffusion coefficient, the radius $R_{\max}(0,t)$ is increasing faster in the brownian case. However the differences

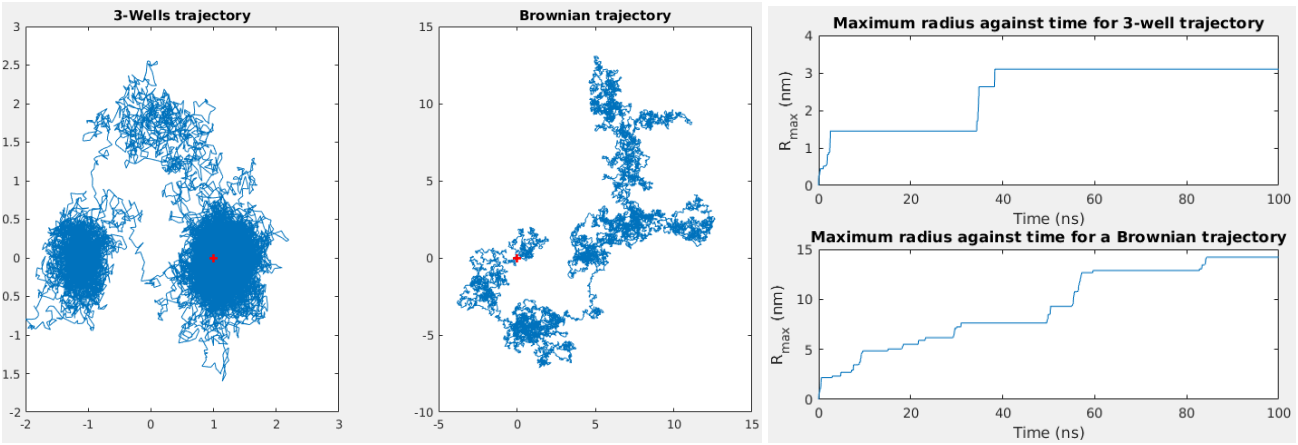


Fig. 1. Left: Three wells trajectory and standard Brownian motion (starting points in red). Right: Evolution od of $R_{\max}$ for both three wells and Brownian case.

between both trajectories appear much more clearly if one compares the evolution of the κ values for both trajectories (see Fig. 2)

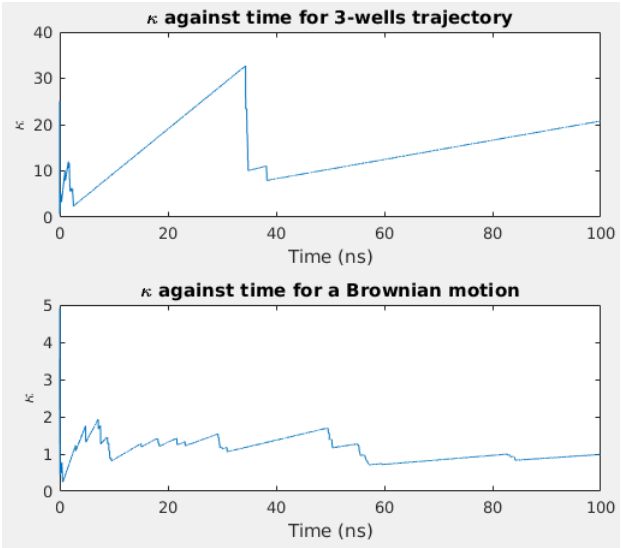


Fig. 2. Profiles of κ in both three-wells case and Brownian case.

As predicted the $t \to \kappa(0,t)$ value stays small (below 2) along the full trajectory in the brownian case whereas it displays clear sharp drops from after reaching linearly local extremum well above 10 in the three well case.

3 *κ-segmentation* algorithm

Following upon this, we built an algorithm providing an automatic segmentation of any MD trajectory, based on the lap number. First, we sum up every possible lap number in a matrix $\kappa(s,t)_{s\leq t}$, within which the algorithm will travel by vertical stripes (see Fig. 3). For a given vertical strip, the width corresponds to a subset of possible candidate conformations at the center of a cluster. Inside a given strip, the analysis is done in a window W traveling along the vertical strip with a fixed stride. Within a given window, the maximum value $\kappa_* = \kappa(s_*, t_*)$, where $[s_*, t_*]$ defined a possible segment corresponding to a portion of the trajectory starting near the center of a cluster at time s_* called hereafter *acces time to the center* – indeed centering that point induces a lower $R_{\max}$ and a greater κ value – and exiting the cluster at time t_* (*exit time*).

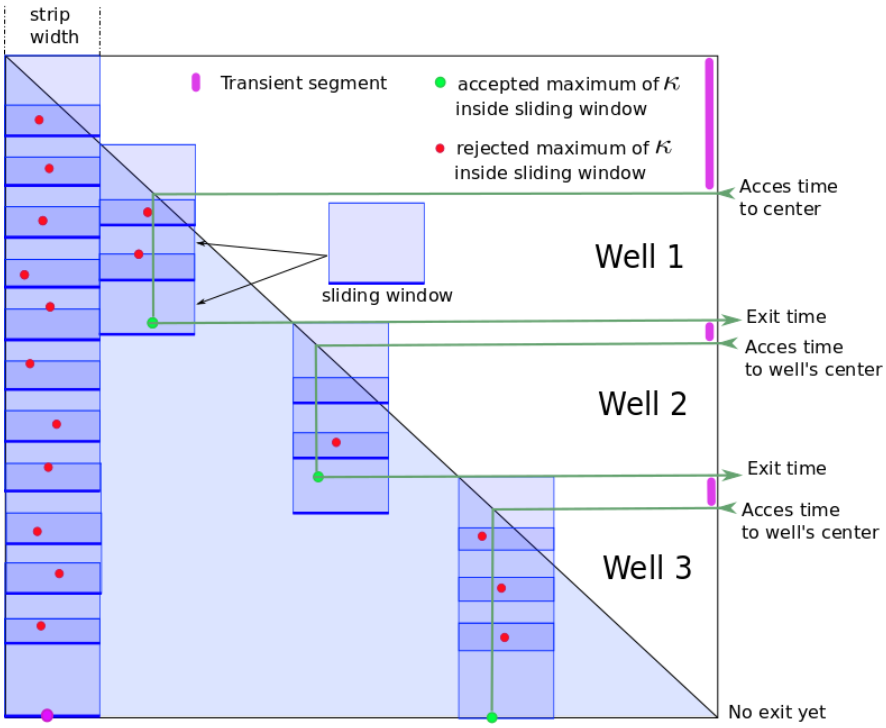


Fig. 3. *κ-segmentation* analysis of the κ matrix. The horizontal axis corresponds to arbitrary starting time s and the vertical axis to arbitrary exit time t with increasing values from top to bottom ($(0,0)$ is at the top left corner)

Then a decision is taken to accept or reject the segment $[s_*\ t_*]$ as a real cluster according the several acceptance test. Briefly we consider two main tests. The first one is to check for $\kappa_* \geq \kappa_0$ where the threshold κ_0 is chosen to prevent against false alarm and spurious detections. The second one is to test if $\inf_{t\in[t_*,t_*+\Delta_0]} R_{\max}(s_*,t)/R_{\max}(s_*,t_*) \geq \rho_0$ ($\rho_0 = 3/4$ in our experiments), to check that the trajectory is not returning too close to the center after exit. The main hyperparameters on the algorithm are then the sliding window size (w_0, h_0), the threshold κ_0 and the test time after exit Δ_0 that have to be chosen of the order of R_0^2/D_0 where R_0 is the expected size of the clusters and D_0 the measured or expected diffusion constant.

4 Applications

4.1 Brownian motion

Here we present the results obtained for the Brownian motion. The lap number matrix shows that κ cannot increase because of a constantly growing $R_{\max}$. As a result, and thanks to the incorporation

of wells rejection criteria, the algorithm only detects transient segments, and do not pronounce any decision concerning the last segment, as it considers its investigation as ongoing.

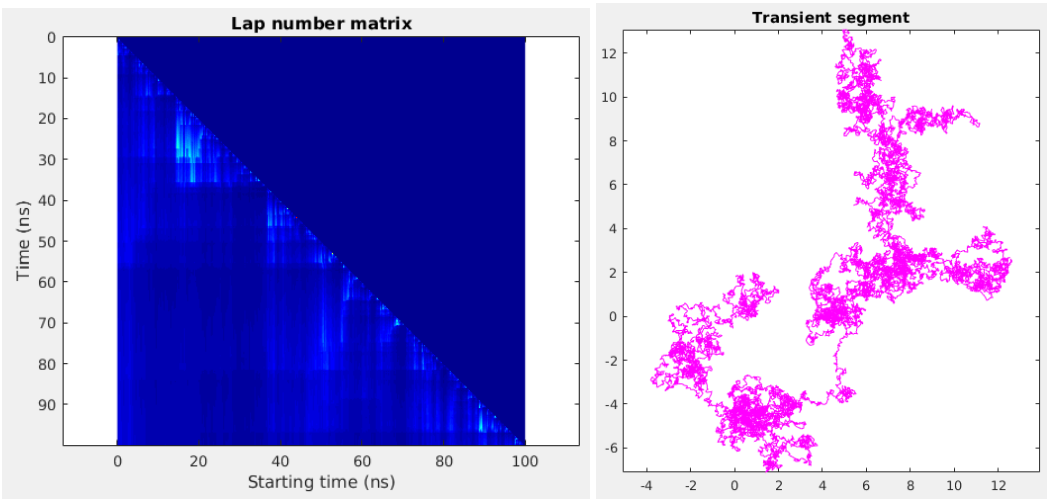


Fig. 4. (Left) lap number matrix. This matrix is read from left to right (shifting the starting point), and from top to bottom (reading the whole trajectory from the starting point). The lap number level is described by the heat intensity. (Right) *κ-segmentation* with transient segments shown in magenta.

4.2 Three wells trajectory

Here we present the results obtained for the three-wells case. The lap number matrix clearly reveals local maximas of κ. The algorithm performed on this trajectory recognize the relevant wells centered in (-1,0) and (1,0), but also points out the third one centered in (0,5/3) as irrelevant, identifying it as a transient segment.

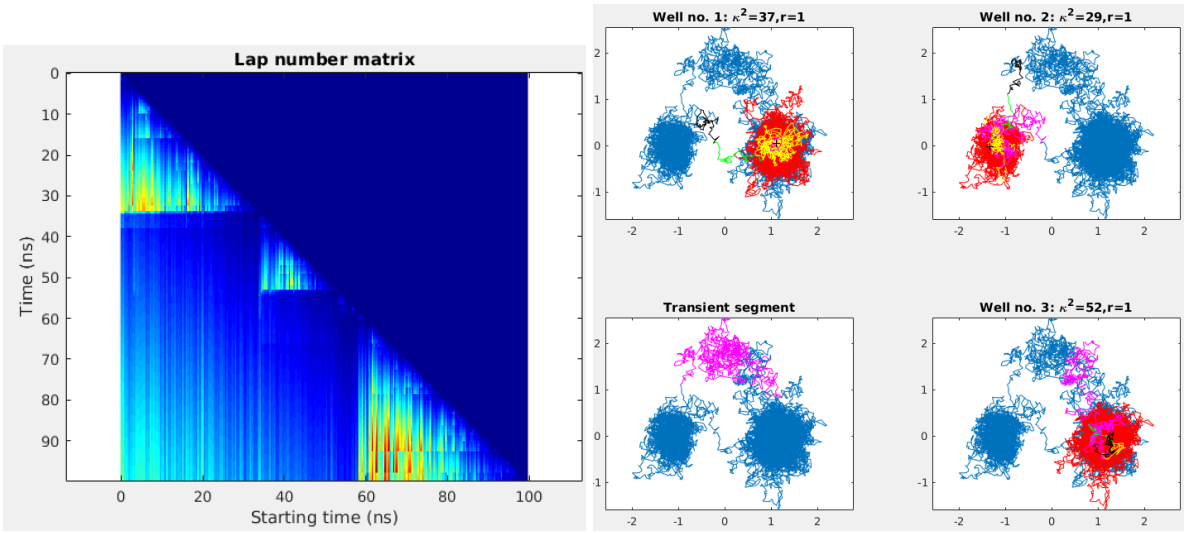


Fig. 5. (Left) lap number matrix. (Right) *κ-segmentation* with transient segments shown in magenta, segments in well are shown in red, sets of initial conformations before and after center are in magenta and in yellow respectively, segment before and after exit are in green and black respectively.

4.3 Clustering of VKORC1 MD simulations

This algorithm has been successfully applied to real MD simulation data generated with GROMACS (AMBER force-field). These MD trajectories were generated for VKORC1 [3], the membrane protein involved in vitamin K recycling that is mandatory for essential physiological processes, such as coagulation, calcium homeostasis, energy metabolism, signal transduction and cell development. Two 5 *ps*-sampled of 1 μs MD trajectories were produced starting from a structural model of VKORC1 in which all cysteine residues were assigned to be protonated (Fig. 5.). We suggested that MD simulation

of such highly flexible model will generate a large conformational space that can be used to predict different enzymatic states of VKORC1 observing upon its enzymatic cycle.

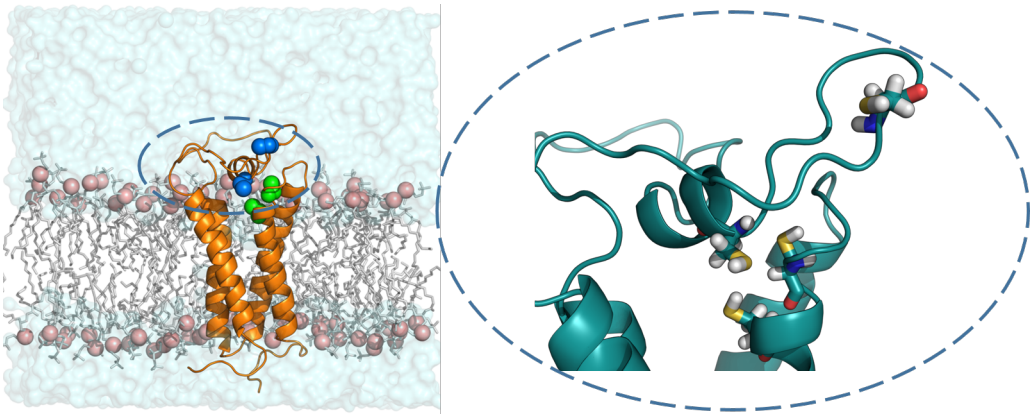


Fig. 6. (Left) VKORC1 protein inserted into membrane (grey, rose) and surrounded by water molecules (blue light). (Right) Viewing of the highly flexible region (contoured) of VKORC1. Protein is shown as cartoon with cysteine residues as balls or as sticks.

In the following, we demonstrate the *κ-segmentation* of the one of these 5 *ps-* sampled 1 μs VKORC1 trajectories. The segmentation was performed after a 2D-projection of the trajectory with a PCA, and after the removal of both lower extremities of the protein located outside the membrane and inducing noise.

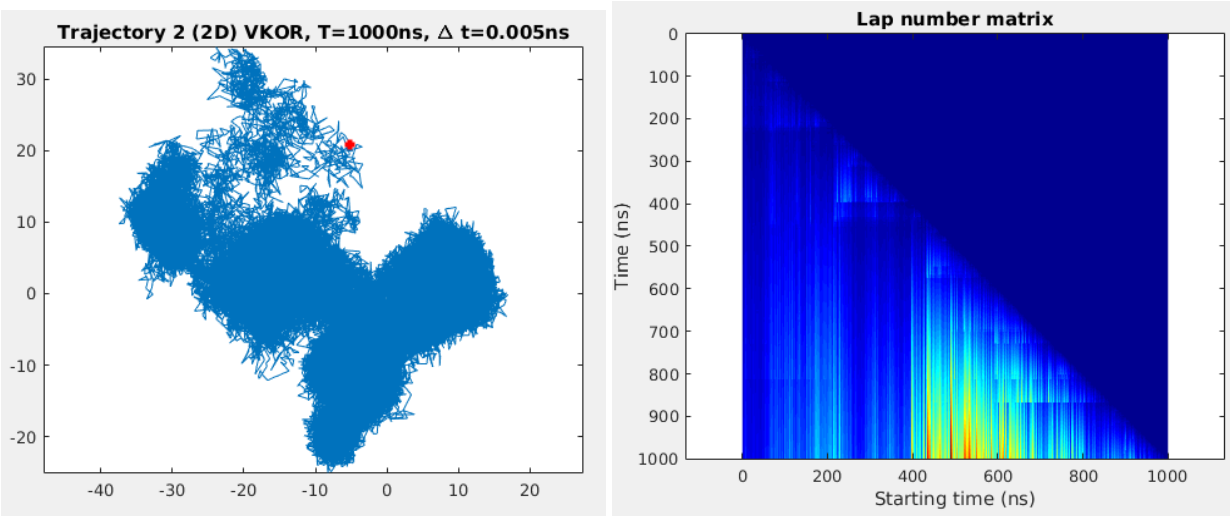


Fig. 7. (Left) 2D-PCA of VKORC1 trajectory. (Right) Lap number matrix.

According to this plot, we can already assume a final well being far deeper than the previous ones, which present cold exit times in terms of lap number, due to the normalization of the plot.

The *κ-segmentation* of this trajectory is presented in Fig. 7. First we observe the very high value of the lap number for the last well ($\sim$ 17000), meaning that the well is very deep. Furthermore, the comparison of wells no. 2 and no. 5 is a good example of a well being deeper than another one, while presenting a lower exit time. In order to obtain reasonably good center points, we had to fix a very large strip width to explore the lap number matrix, $\sim 200 \times \delta t$, where δt represents the time needed by the process to travel a radius distance by diffusion only. If we look closely to the identified wells, we note that the process seems to be trapped in a subset of the well for a short time corresponding to the strip width ($\sim 10\ ns$) before visiting the entire well. This symptom reveals that each well potentially presents several sub-wells, meaning that the algorithm could perform a multi-scale analysis.

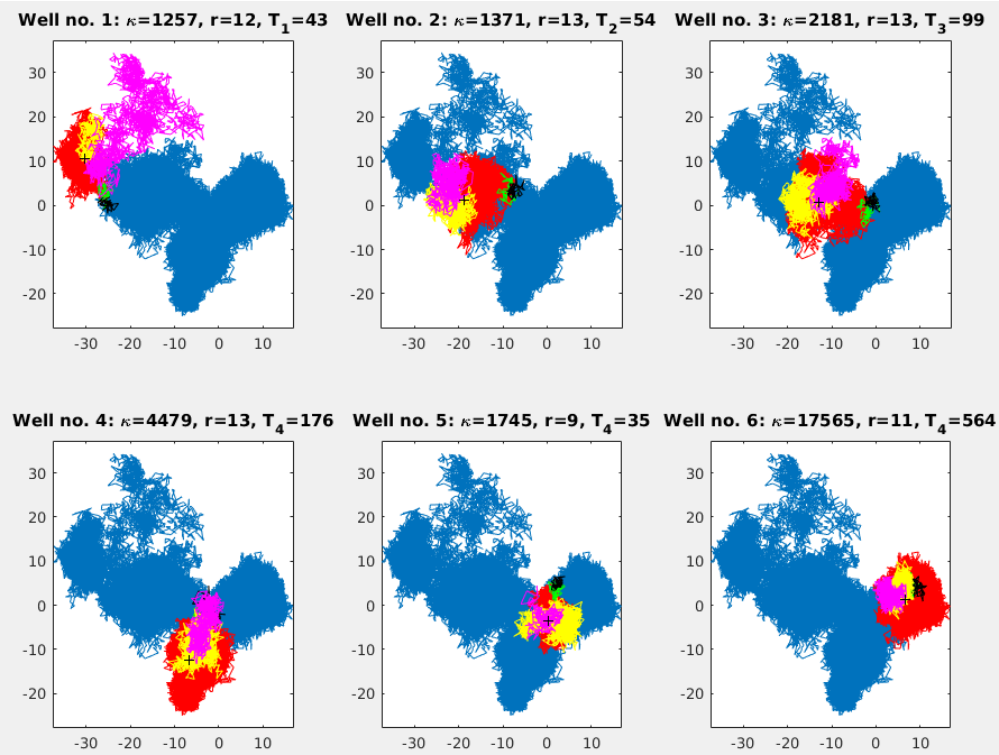


Fig. 8. *κ-segmentation* of the 1 μs MD trajectories of VKORC1 with *κ-segmentation*. Segments in well are shown in red, sets of initial conformations before and after center are in magenta and in yellow respectively, segment before and after exit are in green and black respectively. Each well is associated with its lap number κ, maximum radius r (nm) and exit time T_i (ns).

In the case where wells are overlapping each other, one may use a 3D-PCA of the trajectory, and visualize these wells based on the relevant frames identified with the 2D *κ-segmentation*. For VKORC1, the comparison of wells no. 2 and no. 3 shows an absence of intersection in 3D (Fig. 8). This is remarkable as the algorithm demonstrates the ability to highlight, from 2D dynamics information only, a disjoint phenomenon in higher dimension.

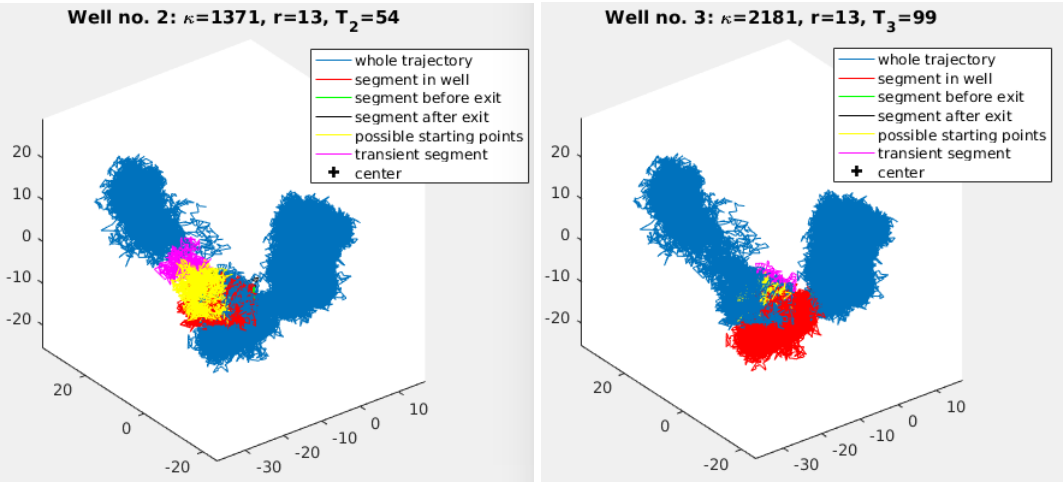


Fig. 9. 3D representation of wells no. 2 and 3 identified with the 2D-*κ-segmentation* of VKORC1 trajectory. Each well is associated with its lap number κ, maximum radius r (nm) and exit time T_i (ns)

4.4 VKORC1 clusters and their interpretation

Application of *κ-segmentation* to each MD trajectory of VKORC1 produced a set of six clusters with different population. As the population of a cluster is associated with its lap number κ, we used this value for the clusters comparison. In both sets the most populated cluster was observed at the end

of the simulation. We suggest that the last cluster may be composed of conformations representing the fully equilibrated protein's state.

To interpret the content of each cluster localized by *κ-segmentation*, a pair of VKORC1 conformations located within the same well, one in the center of a well and the other at exit time from the well, were superimposed. In the both MD replica, each pair of conformations within a well demonstrates a modest but pronounced difference. In contrast, if we compare the distinct wells, the respective conformations are highly differed in a folding and in a position of the flexible regions of VKORC1 - the L-loop and the N- and C-terminals. At the beginning of MD simulations (the 1-st and the 2-nd replica), the L-loop structure displays an unique short alpha-helix (L-helix) situated in the middle of the extended coiled structure (Fig. 10, W1-W3, Top and Bottom).

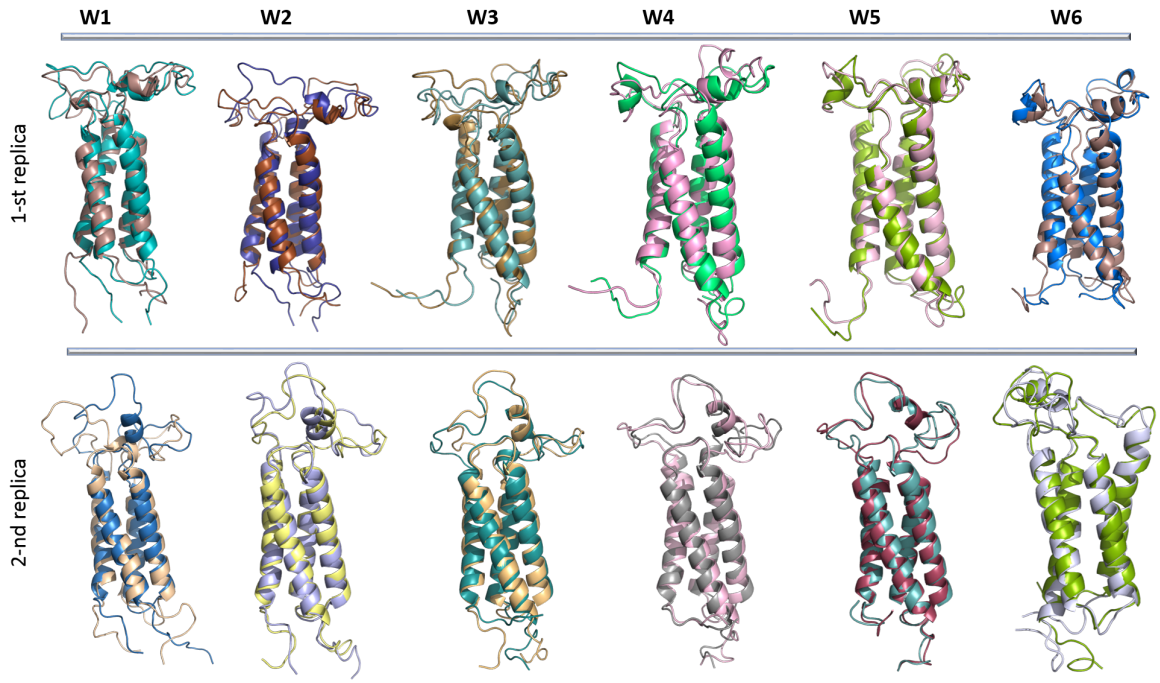


Fig. 10. Superimposition of VKORC1 conformations corresponding to the well's center (W1, turquoise/camel; W2, orange/grey; W3, brown/brown light; W4, rose/grey; W5, pink/red dark; W6, brown/green) and at exit time (W1, brown/blue; W2, violet/yellow; W3, blue/deepteal; W4, green/pink; W5, green light/turquoise; W6, blue/white) illustrated for each well of the 1-st (Top) and the 2-nd (Bottom) (1-st/2-nd) replica of MD simulation respectively. Protein is represented as cartoon with PyMol.

Analyzing the extended MD simulations, we observed that during the 1-st replica, (W4-W6), the L-loop adopts a stably folded structure that shows, additionally to the L-helix, formation of two novel 310-helices leading to the better stability and compactness of VKORC1 at the end of simulation (Fig. 10, W3-W6 at the Top). In the 2-nd replica, the unique L-helix of L-loop is highly conserved in all wells but its position relative to the transmembrane region (TMR) is significantly changed from 'a closed -TMR)' to 'a distal-TMR'. As decoded by *κ-segmentation*, two replica represent two different ways of evolution of the initial structure (at t=0 ns) of VKORC1 during MD simulation. The reconstitution of these two processes in terms of structure/conformation is illustrated in Fig. 11. Two extended MD simulations of the initial VKORC1 conformation lead to stabilization of two distinct enzymatic states of VKORC1 observed during its catalytic cycle. The cluster-based characterization, obtained with *κ-segmentation*, allows to follow interactively the simulation process and to describe finely the conformational and structural change in a highly flexible protein over the simulation of its dynamics as was illustrated for VKORC1.

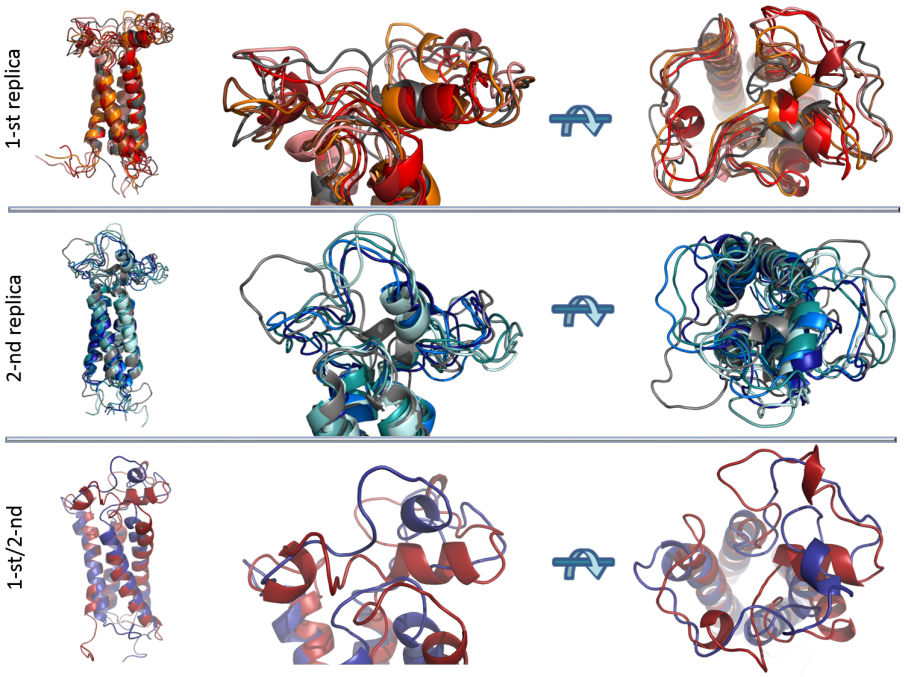


Fig. 11. Superimposition of VKORC1 conformations picked at the center of each well.Top and Middle : Two replica (1-st/2-nd) of MD simulation. Color code: W1 (grey/grey), W2 (brown/lightcyan), W3 (salmon/cyan), W4 (orange/deepteal), W5 (red/blue), W6 (red dark/ blue dark). Bottom: Superimposition of conformations picked at W6 (red/blue) of both MD trajectories (1-st/2-nd). Superimposed conformations (left) were zoomed on the L-loop and represented in two orthogonal projections.

5 Conclusion and perspectives

The algorithm described here allows the user to perform a dynamic based segmentation of MD-trajectories driven by the analysis of the lap number κ to quantify the depth of a well (metastable state). The method is robust to false alarm and produces a sequence of wells coming with descriptive metrics (entrance and exit time, center of the well, lap number) allowing comparison between wells. Using the metrics, we can for instance characterize the metastable states, decide if the simulation is long enough to reach a relevant equilibrated state of protein or is worth being extended. Moreover, the *κ-segmentation* could provide a hierarchical cluster representation containing information on metastables states at different levels, which can further help the user to characterize the architecture of the free energy landscapes. The κ criteria provides an absolute description of wells depths, whereas methods such as spectral clustering or Jarvis-Patrick clustering, due to their metrics-based analysis ignoring the dynamic aspects, do not allow the user to evaluate the relevance of a well. However, we note that the *κ-segmentation* method needs additional development such as graphic interface that can help identify a structural content of the wells. Also, consideration of multi-component systems, such as the protein complexes is required.

Acknowledgements

We warmly thank Myriam Hanna (CMLA, ENS Paris-Saclay) for technical help.

References

[1] Jun-hui Peng, Wei Wang, Ye-qing Yu, Han-lin Gu, and Xuhui Huang. Clustering algorithms to analyze molecular dynamics simulation trajectories for complex chemical and biological systems. *Chinese Journal of Chemical Physics*, 31(4):404–420, 2018.

[2] Boaz Nadler, Stéphane Lafon, Ronald R Coifman, and Ioannis G Kevrekidis. Diffusion maps, spectral clustering and reaction coordinates of dynamical systems. *Applied and Computational Harmonic Analysis*, 21(1):113–127, 2006.

[3] Nolan Chatron, B Chalmond, A Trouvé, Etienne Benoit, H Caruel, Viginie Lattard, and L Tchertanov. Identification of the functional states of human vitamin k epoxide reductase from molecular dynamics simulations. *RSC Advances*, 7(82):52071–52090, 2017.

Reference-guided genome assembly in metagenomic samples

Cervin Guyomar[1,2], Wesley Delage[1], Fabrice Legeai[1,3], Christophe Mougel[3], Jean-Christophe Simon[3] and Claire Lemaitre[1]

[1] Univ Rennes, Inria, CNRS, IRISA, F-35000 Rennes, France
[2] iDiv, Deutscher Pl. 5E, 04103 Leipzig, Germany
[3] INRA EGI, F-35000 Rennes, France

Corresponding author: cervin.guyomar@idiv.de

Abstract *Assembling genomes from metagenomic data is a challenging task, because of both the many species coexisting in the samples and the polymorphism within these species. Most approaches consist in a complete assembly of the metagenome into contigs, that can then be binned into taxonomic units. On the opposite, we present in that work a targeted assembly approach in two steps. First, taking advantage of a potentially distant reference genome, a subset of the metagenomic reads is assembled into specific contigs. Then, using an enhanced version of the MindTheGap local assembly algorithm, this first draft assembly is completed using the whole metagenomic readset in a de novo manner. The resulting assembly can be output as a genome graph, allowing to distinguish different strains with potential structural variants coexisting in the sample. MindTheGap was applied to 32 pea aphid re-sequencing samples in order to recover the genome sequence of its obligatory bacterial symbiont, Buchnera aphidicola. It was able to return high quality assemblies (one contig assembly in 90% of the samples), even when using increasingly distant reference genomes, and to retrieve large structural variations in the samples. Due to its targeted approach, it outperformed standard metagenomic assemblers in terms of both time and assembly quality. As such, it appears as a promising approach for single genome assembly from metagenomic data.*

Availability: *https://github.com/GATB/MindTheGap*

Keywords Genome assembly, metagenomics, reference-guided, short reads

1 Introduction

The advances of molecular techniques revealed the importance of microorganisms in every ecosystem. In particular, whole-genome metagenomic sequencing makes it possible to understand the full functional potential of microbial communities by accessing the whole genomic sequence of both culturable and unculturable microbes. However, extracting relevant information from complex metagenomic datasets is a challenging task. Current metagenomic datasets are a mixture of short reads originating from different species. Thus, reconstructing genomes from metagenomic data requires two steps : the assembly of reads into longer sequences, and the partitioning of sequences based on their taxonomic origin.

Metagenomic assembly consists in forming contigs prior to the taxonomic binning of sequences. Many recent software are devoted to this task [1,2,3]. However, because of the high complexity of such data, *de novo* assembling contigs from metagenomic reads is challenging and comes with a high computational cost. Metagenomic assemblies are very fragmented because of homologous regions between microbial species and polymorphism within the species [4].

An alternative to this approach would be to partition in a first step the metagenomic reads into subsets assigned to different species. Binning methods relying on the nucleotidic composition of reads cannot be applied to the current Illumina reads because of their short length [5]. Alternatively, it is possible to select reads by reference-based approaches, but these approaches struggle to classify reads from badly known species, and hardly scale up to large datasets when based on alignment methods [6]. A relevant strategy to assemble a given genome from metagenomic data is to map reads against the closest available reference genome to assemble new contigs. The quality of the assembly is therefore highly dependent on the evolutionary distance with the reference genome. In particular, any region absent or too divergent from the reference genome will be missed. This enables nonetheless

the targetted assembly of a genome of interest within a community, which is for instance relevant for the study of key players of host-symbiont relationships or the discovery of new pathogenic strains of known microbes. In these use cases, functional, structural or phylogenetic genomic analyses require the assembly of a new genome of interest from metagenomic data. In that context, neither *de novo* metagenomic assembly nor assembly from reads selected by reference alignment are able to return assemblies of good quality. Nonetheless, it seems possible to use the best of these two strategies, by selecting reads from regions of homology with a related reference genome, and using *de novo* assembly to reconstruct the missing regions.

Several tools, such as MITObim [7], LOCAS [8], Pilon [9] or IMR/DENOM [10], were designed following this idea, combining reference alignment and *de novo* assembly. However, all of them show some limitations because of which they are not adapted to metagenomic data. They are either not scaling up with these large datasets (MITObim, LOCAS), unable to deal with large structural variants (IMR/DENOM, Pilon, LOCAS) or to return coexisting variants (LOCAS, IMR/DENOM).

In this work, we present a solution for the assembly of a genome of interest from metagenomic data, in a reference-based manner. This method can recover large regions absent from the given reference genome, makes no assumption on the ordering or direction of regions homologous with the reference and is able to return several different solutions reflecting the metagenomic diversity inside the sample. The method is based on two main steps, a reference based recruiting and assembly of metagenomic reads, followed by a targetted assembly, filling the gaps between the contigs assembled beforehand.

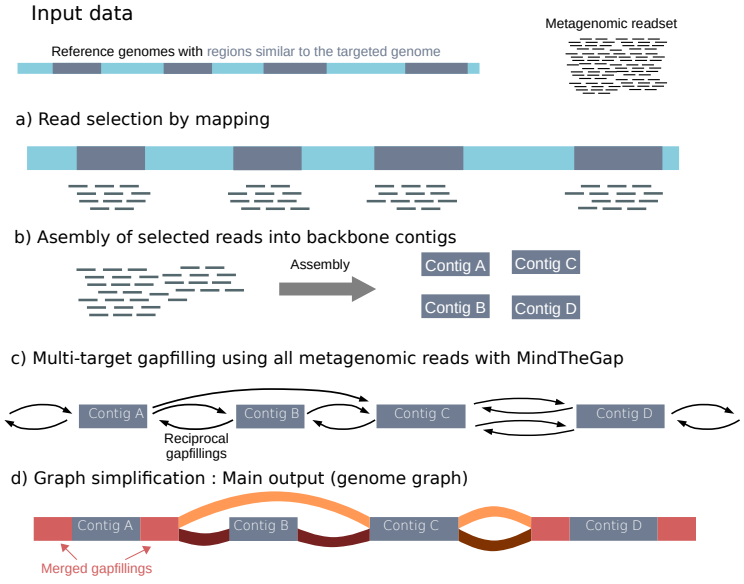
We applied this method to reconstruct genomes from several metagenomic samples of the pea aphid *Acyrthosiphon pisum*. Focusing on the primary endosymbiont *Buchnera aphidicola*, we demonstrated the ability of *MindTheGap* to assemble complete bacterial genomes in a single contig using a remote genome as a primer, even when structural variability is present.

2 Material and Methods

2.1 Targeted assembly for metagenomic data

Strategy overview The method described in this work relies on a two-step pipeline, described in Figure 1.

Fig. 1. Overview of the *MindTheGap* reference-guided assembly pipeline



The first step uses a given reference genome to build an incomplete but trustworthy assembly, matching with the conserved regions of the genome. The second step uses the whole set of metagenomic reads to extend the previously assembled contigs and form a complete assembly, without any *a priori* on the order and orientation of contigs. The result of the pipeline is a genome graph encompassing the stuctural diversity detected on the assembled genome. This graph can be exploited by extracting contigs, or paths of the graph that represent different strains.

2.1.1 Assembly of backbone contigs The first step requires a metagenomic readset and a reference genome, and returns contigs that are assembled using reads mapped on the reference. All metagenomic reads are mapped against the reference genome using BWA MEM [11], and the mapped reads are kept and *de novo* assembled using the Minia [12] assembler. Although any assembler can be used in this step, we use Minia [12] for its low memory footprint, and its assembly algorithm similar to the one used in the second step of the method. The goal of this step is to generate high quality contigs, that can reliably be used for the upcoming gapfilling. To ensure this, we set up Minia with more stringent parameters than for an usual assembly task and only contigs longer than a user-defined threshold (500 bp by default) are kept.

2.1.2 Parallel gapfilling with MindTheGap The essential step of the pipeline is the gapfilling between backbone contigs, which enables the assembly of regions absent from the reference genome. This is made possible by a targeted assembly of the whole readset using the previously assembled contigs as primers. This step does not require the ordering of contigs, since all possible combinations are tested during gapfilling. As a result, structural variants can be detected, either compared to the reference genome or within the sample.

This step is based on a module of the software *MindTheGap*, originally developed for the detection and assembly of insertion events [13]. The *fill* module of *MindTheGap* performs a local assembly for each pair of breakpoint event kmers, resulting in one or several insertion sequences.

In this work, we took advantage of this module of *MindTheGap* and adapted it to the problem of reference-guided assembly. It has been modified to make possible the gapfilling between a seed kmer and **multiple** target kmers, enabling the "all versus all" gapfilling within a set of contigs with only a linear increase of the runtime (compared to a quadratic increase for a naive "all versus all" gapfilling). The resulting algorithm is presented in Figure 2. A seed kmer is extracted at the end of each contig and its reverse-complement, resulting in a set of $2n$ kmers for n contigs. Similarly, a set of $2n$ target kmers is extracted at the beginning of each contig and its reverse-complement. For each seed kmer, a contig graph is created by starting from the seed kmer and performing a breadth first traversal of the *De Bruijn* graph representation of the whole readset. Contigs are consensus sequences returned by removing graph motifs such as bubbles (SNPs) and tip-ends (errors). In the contig graph, contigs are nodes, and edges represent the existence of a $k-1$ nucleotide overlap between two contigs. The creation of the contig graph is similar to the one used in *Minia* [12]. The traversal is stopped when the graph becomes too large (total assembled nucleotides) or too complex (number of contigs), following user-defined parameters. Importantly, if one of the target kmers is found during the contig graph construction, that contig is not extended further, avoiding redundant contig assembly, and saving time and memory. After the contig graph has been built, target kmers are searched within this contig graph, and gapfilling sequences are built, by retraversing the contig graph from the seed kmer to contigs containing a given target kmer. For each seed-target couple, if several solutions are returned, redundant solutions above a 95% identity threshold are removed. Thanks to this multi-target version of the algorithm, only $2n$ contig graph constructions are necessary to search possible sequences between all pairs of contigs, instead of n^2 with the naive approach.

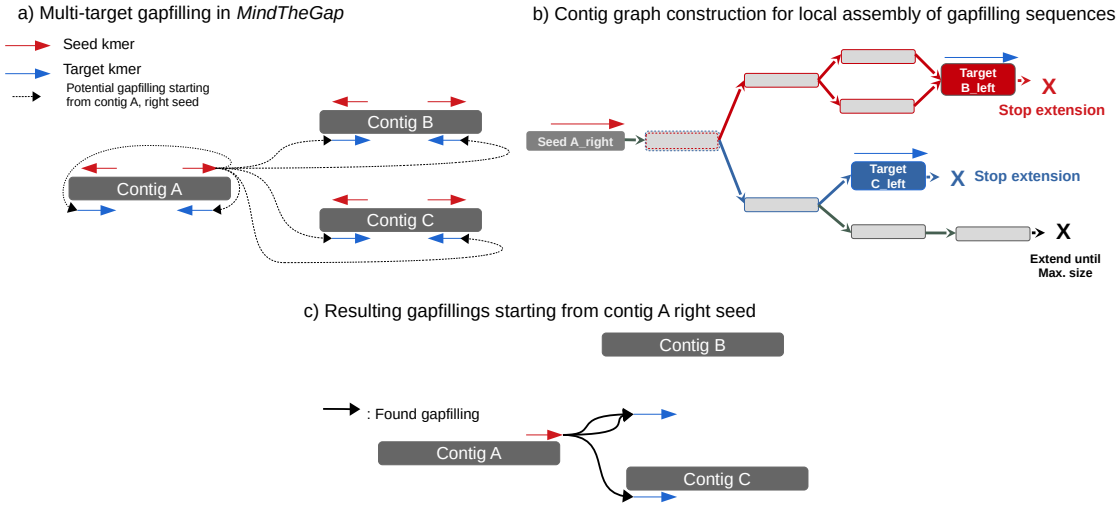
The whole process is parallelized by dispatching the $2n$ starting kmers to different threads. The main output is a genome graph in the GFA format (Graphical Fragment Assembly, https://github.com/GFA-spec/GFA-spec), giving the overlap relationships between contigs an their gapfillings.

2.1.3 Graph simplification and visualisation In order to return a standard fasta assembly, the genome assembly has to be processed. The complexity of the graph is reduced on several steps using a post-treatment program.

First, it is likely that two contigs are linked in the graph by two gapfillings with reverse-complement sequences, one starting from the left contig and the other one starting from the right contig. Such reciprocal links are removed, when their sequence identity is over a 95% threshold.

Secondly, when several gapfilling sequences start (or end) from the same seed, it is possible that a subset of them have an identical prefix (suffix) and start to diverge after a potential large distance to the seed. This results in redundant sequences in the graph. A node merging algorithm is applied, in order to return contigs that do not share large identical subsequences (prefix or suffix). Sets of

Fig. 2. Gapfilling a set of contigs using *MindTheGap* fill module.
a) Seed and target kmers are extracted from the 3 input contigs, resulting in 2 sets of 6 kmers, seed (red) and target (blue) ones. b) A graph of contigs is built starting from the right seed kmer of contig A. Extension is stopped when a target kmer of another contig if encountered, or a maximum assembly size is reached. c) This results in 3 gapfilling sequences starting from contig A right seed, 2 gapfilling sequences joining contig B, and one contig C.



sequences sharing the same 100 first nucleotides are built. Within each set, the sequences are then compared to find the first divergence between all sequences. A new node is added to the graph, containing the repeated portion of the sequences, and repeated nodes are shortened accordingly. This process is applied iteratively to every node, including the newly created nodes, for which a subset of neighbors may still show identical sequences.

Finally, simple linear paths in the graph are merged, nodes whose length is lower than 500 bp are removed, and highly branching nodes (connected to more than 5 contigs) are cut. The resulting graph is a good representation of the *MindTheGap* assembly, and nodes can be extracted to be used as regular contigs.

After the simplification process, the graph may not be a linear sequence because of intra-sample polymorphism or assembly uncertainties. The final assembly can be generated either by manual inspection of the graph using the *Bandage* software [14], or by enumerating all possible paths within the graph.

2.1.4 Implementation and availability *MindTheGap* has been officially released in version 2.1, enabling the so-called "contig mode" for reference-guided assembly (https://github.com/GATB/MindTheGap). *MindTheGap* is written in C++ using the GATB library [15] (https://github.com/GATB/gatb-core). The GATB library provides algorithms for the analysis of NGS datasets with high performances and a low memory footprint. The graph simplification is performed using Python scripts, available on the *MindTheGap* repository. A complete pipeline including mapping, assembly and gapfilling is also available as a Python script distributed along with *MindTheGap*.

2.2 Application to pea aphid metagenomic datasets

In this study, we applied *MindTheGap* assembly pipeline to the assembly of the obligatory bacterial symbiont of the pea aphid holobiont, *Buchnera aphidicola*. We considered 32 pea aphid resequencing samples of paired end 100bp *Illumina* reads. These datasets have already been studied in a previous work, in which the microbiota of each aphid sample was detailed [16]. The number of reads per dataset is ranging from 65 to 118 million, with an average coverage of 628X for the *Buchnera* genome.

Reference genomes with increasing levels of divergence Reference-guided assembly was performed with 4 distinct reference genomes of *Buchnera aphidicola* with different levels of divergence : 1) *Buchnera aphidicola* from A pisum (LSR1 accession), hereafter called *Buchnera LSR1*, which is the closest available assembled genome; 2) *Buchnera* from *Myzus persicae*; 3) *Buchnera* from *Uroleucon*

ambrosiae the most divergent reference analyzed ; and 4) a synthetic genome obtained by deleting 116.4 Kb of sequences from *Buchnera LSR1*. The synthetic rearranged LSR1 genome was generated by applying 20 deletions, whose size ranged from 300 bp. to 20 kbp. The levels of divergence are supported by phylogenetic studies [17] and genome alignment. *Buchnera LSR1* was aligned on the *Myzus persicae* with a 93% coverage, and to *Uroleucon ambrosiae* witha 87% coverage, with a genome identity of 80% on the aligned regions.

Inclusion of simulated structural variations To assess the ability of *MindTheGap* to recover structural variations in samples with strain diversity, we created a synthetic pea aphid sample by adding to a randomly chosen real sample, a subset of simulated reads from the previously described rearranged genome (with 20 deletions). 50X coverage of reads were simulated with *wgsim* of the Samtools suite.

MindTheGap assembly pipeline parameters *MindTheGap* was used in version 2.2.0, with the same set of parameters for all samples and reference genomes. For the assembly step, a *kmer* size of 61 was chosen, along with a solidity threshold of 10, and a minimum contig length of 400 bp. The gapfilling step was performed using a k value of 51, and a solidity threshold of 5.

Comparison with other approaches The results were compared to those of a usual approach to assemble a particular genome from metagenomic data. A complete *de novo* assembly was performed for each sample using MegaHit [3] and *Buchnera* contigs were selected by a Blast alignment against the genome of *Buchnera aphidicola APS*. Only contigs with at least 50% of the length covered by Blast hits with e-value smaller than 10^{-5} were kept.

The quality of each assembly was assessed using Quast [18] and the reference genome of *Buchnera aphidicola APS* from *A. pisum*. Similarly to what was done with *MindTheGap*, we did not include contigs smaller than 1 Kb, mainly associated with plasmid sequences.

3 Results

3.1 Single chromosome assembly of *Buchnera aphidicola* from metagenomic data

MindTheGap assembly pipeline was applied on 32 pea aphid resequencing samples [16] to assemble its bacterial obligatory symbiont *Buchnera aphidicola* (640 Kb). These are metagenomic samples comprising the insect host genome together with is microbial symbiotic communities. More than 90% of the reads originate from the insect host, and are not relevant when focusing on symbiont genomes. This particular fact motivates the choice of a targeted assembly technique, which does not require to assemble all the pea aphid reads.

In order to assess the robustness of the approach with respect to the level of divergence of the reference genome, four different genomes of *Buchnera aphidicola* of increasing divergence were used as a guide for the assembly, and the resulting contigs were compared to the closest reference available as a validation.

A summary of the assemblies obtained using the different reference genomes is shown in Table 1. When using either *A. pisum (LSR1)* or *M. persicae* reference genomes, most samples were assembled in a single contig whose length is very close to the target length (Less than 1% length divergence, or 6 kb). 91% of samples were assembled in a single complete contig. Using *Buchnera* from *Uroleucon* as a guide returns less complete assemblies, with only 65% of the samples that were fully assembled. This is due to its greater evolutionary distance to the genome to assemble, This greater distance is particularly well exemplified when looking at the relative contributions between the two steps of the pipeline, mapping based assembly and *de novo* gapfilling. Only an average of 6.92% of the target genome is assembled after the first step when using *Uroleucon's Buchnera*, whereas this fraction is of 47.6 % for *Myzus* and 99.9% for *Acyrthosiphon*.

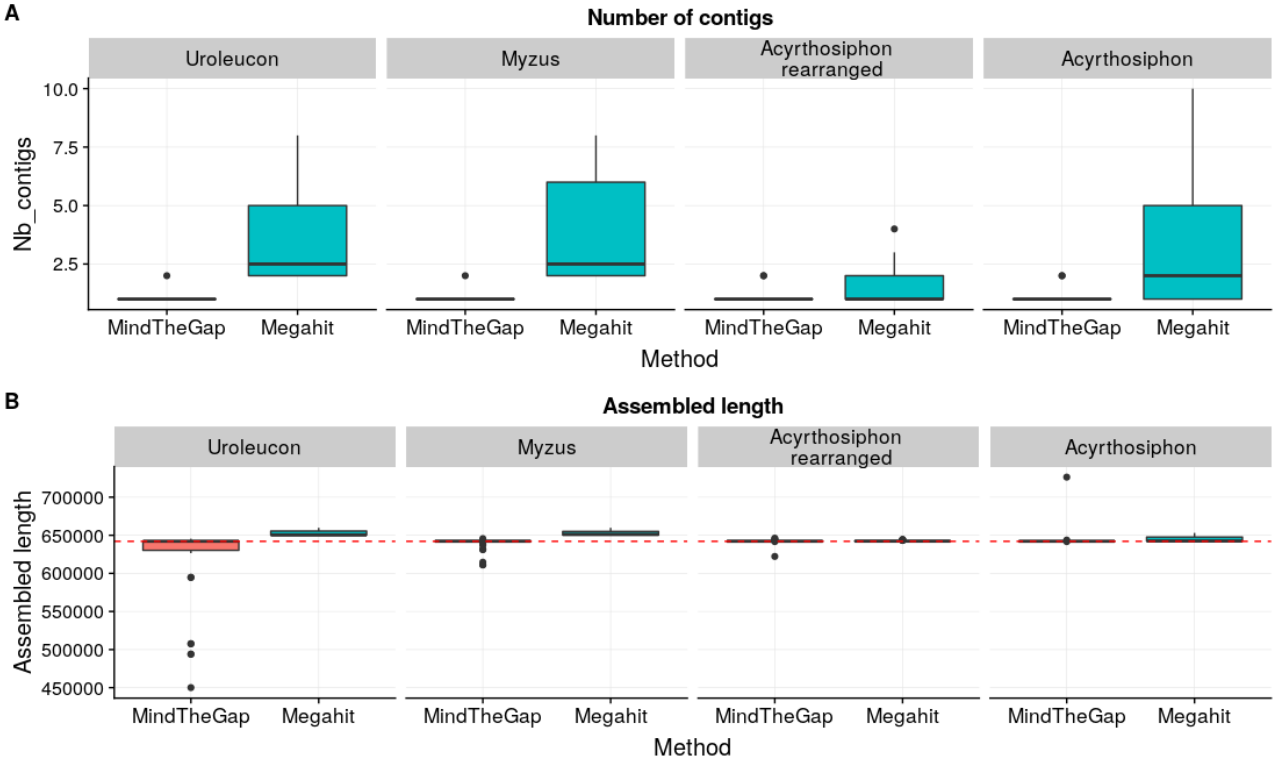
When using a rearranged genome missing several large sequences (totaling 116.4 Kb), most samples were also assembled into a single contig and all the missing regions were fully recovered. Although the complete genome length was recovered, less circular contigs were returned compared to other reference genomes.

Comparison with a classical metagenomic assembly The assemblies performed by *MindTheGap* were compared to those of an alternative stratergy, consisting in a *de novo* assembly using *MegaHit* [3] followed by a selection of contigs using a reference genome.

	Buchnera *Uroleucon ambrosiae*	Buchnera *Myzus persicae*	Buchnera rearranged	Buchnera *Acyrthosiphon pisum*
Circular complete assemblies	20	22	17	22
Linear complete assemblies	1	5	12	7
2 contig complete assemblies	1	2	2	2
Incomplete / erroneous assemblies	9	3	1	1

Tab. 1. Overview of assembly results with four different *Buchnera* reference genomes used as guide, from the closest relative at the right, to the most distant at the left. A complete assembly has a size with no more than 1% variation compared to the reference genome *Buchnera* LSR1.

Fig. 3. Number of contigs (A) and assembly length (B) using four different *Buchnera* genomes as assembly guide. The expected genome length (*Buchnera* LSR1) is shown as a red dotted line.



For most samples, *MindTheGap* outperforms the metagenomic assembly by returning assemblies with less contigs, and a total length closer to the expected genome size. Reference-guided assembly enables a one-contig assembly in most cases (90%), whereas *MegaHit* outputs a single contig for only 28% of samples. The average assembly size for *MegaHit* exceeds the expected genome length. An explanation for this could be that highly polymorphic regions may be assembled into distinct contigs by the metagenomic assembler, while *MindTheGap* merges them, or represents them as bubbles in the genome graph.

Importantly, *MindTheGap* is also significantly faster than *MegaHit*. The average runtime of *MindTheGap* assembly pipeline is 95 minutes, which is 5.5 times inferior to *MegaHit* runtime (525 minutes). Indeed, *MegaHit* produces contigs not only for the target organism, but in this case for the insect host *A. pisum* and its secondary symbionts.

3.2 Assembly of large structural variations in a metagenomic context

MindTheGap was applied to a pea aphid sample in which simulated reads from a rearranged *Buchnera* genome were added, simulating the coexistence in a metagenomic dataset of two strains with structural variations. In the resulting genome graph, 17 out of the 20 simulated deletions were fully recovered, with both the deleted and complete versions of the genome assembled. Extracting the longest path from the graph resulted in a one contig 641,531 bp assembly, compared to the 642,011 bp of the *Buchnera* LSR1 genome. Similarly, the shortest path extracted from the graph was 526,448bp long, compared to 525,611 for the deleted simulated genome. The longest structural variations (up to 20 Kb) were all successfully recovered. Only two 500 bp and one 300 bp variations were missing from the graph.

The metagenomic assembly with *MegaHit* of the same readset, followed by a filtering of contigs using the deleted reference genome, resulted in a 38 contigs assembly, with a length of 645,973 bp and a N50 of 44,484 bp. It highlights the difficulty of *de novo* assembly to deal with structural diversity in metagenomic samples.

4 Discussion and conclusion

Starting from the observation that both reference-based assignation and *de novo* assembly are inadequate to study some aspects of the metagenomic diversity, we present in the present work an hybrid method under the term of reference-guided assembly. This method was designed to assemble the genome of a single species of interest and its structural variants from a potentially larege and complex metagenomic dataset. We have shown here that it outperforms both reference-based approaches and *de novo* assemblers. Reference based read assignation is highly dependent on the evolutionary distance of the targeted genome with available references. This was particularly highlighted in this work, where less than 10 % of the genome could be assembled with the reads mapping to the most divergent reference genome used in this analysis. In *de novo* approaches, the assembly is performed prior to contig binning or mapping. This can be described as an *Assembly-first* approach. Here, we present a *Mapping-first* approach, that lightens the computational burden of full *de novo* metagenomic assembly, at the cost of a single genome assembly. To our knowledge, this is the first reference-based assembly approach suitable for metagenomic data.

Beyond the pea aphid complex, *MindTheGap* may also be applied to a wide range of assembly issues. The targeted assembly approach reduces the number of sequences to assemble, and thus simplifies the assembly problem. This approach may therefore be suitable for large and complex communities such as the human microbiome. Here, *MindTheGap* was presented as a complete pipeline from reads to contigs, but the second step of the pipeline can be associated to any other assemblers. In this manner, *MindTheGap* can be used as a finishing tool for previous incomplete assemblies. In a metagenomic context, the gapfilling step may be a way to increase the contiguity of assemblies by joining metagenomic contigs identified by binning methods as coming from the same species.

A valuable feature of *MindTheGap* is to output a genome graph representation instead of a set of unconnected contigs. This is particularly useful to represent the structural diversity of the genomes, which is rarely examined in metagenomic datasets.

Acknowledgements

Computations have been made possible thanks to the resources of the Genouest infrastructure.

References

[1] Sergey Nurk, Dmitry Meleshko, Anton Korobeynikov, and Pavel A. Pevzner. metaSPAdes: a new versatile metagenomic assembler. *Genome Research*, 27(5):824–834, mar 2017.

[2] Yu Peng, Henry C M Leung, S. M. Yiu, and Francis Y L Chin. IDBA-UD: A de novo assembler for single-cell and metagenomic sequencing data with highly uneven depth. *Bioinformatics*, 28(11):1420–1428, jun 2012.

[3] Dinghua Li, Chi-Man Liu, Ruibang Luo, Kunihiko Sadakane, and Tak-Wah Lam. MEGAHIT: an ultra-fast single-node solution for large and complex metagenomics assembly via succinct de Bruijn graph. *Bioinformatics (Oxford, England)*, 31(10):1674–6, may 2015.

[4] Alexander et al Sczyrba. Critical Assessment of Metagenome Interpretation - A benchmark of metagenomics software. *Nature Methods*, 14(11):1063–1071, 2017.

[5] Hanno Teeling, Anke Meyerdierks, Margarete Bauer, Rudolf Amann, and Frank Oliver Glöckner. Application of tetranucleotide frequencies for the assignment of genomic fragments. *Environmental Microbiology*, 6(9):938–947, sep 2004.

[6] Daniel H. Huson, Sina Beier, Isabell Flade, Anna Górska, Mohamed El-Hadidi, Suparna Mitra, Hans Joachim Ruscheweyh, and Rewati Tappu. MEGAN Community Edition - Interactive Exploration and Analysis of Large-Scale Microbiome Sequencing Data. *PLoS Computational Biology*, 12(6):e1004957, 2016.

[7] Christoph Hahn, Lutz Bachmann, and Bastien Chevreux. Reconstructing mitochondrial genomes directly from genomic next-generation sequencing reads - A baiting and iterative mapping approach. *Nucleic Acids Research*, 41(13):e129, jul 2013.

[8] Juliane D. Klein, Stephan Ossowski, Korbinian Schneeberger, Detlef Weigel, and Daniel H. Huson. LOCAS – A Low Coverage Assembly Tool for Resequencing Projects. *PLoS ONE*, 6(8):e23455, aug 2011.

[9] Bruce J. Walker, Thomas Abeel, Terrance Shea, Margaret Priest, Amr Abouelliel, Sharadha Sakthikumar, Christina A. Cuomo, Qiandong Zeng, Jennifer Wortman, Sarah K. Young, and Ashlee M. Earl. Pilon: An integrated tool for comprehensive microbial variant detection and genome assembly improvement. *PLoS ONE*, 9(11):e112963, nov 2014.

[10] Xiangchao Gan, Oliver Stegle, Jonas Behr, Joshua G. Steffen, Philipp Drewe, Katie L. Hildebrand, Rune Lyngsoe, Sebastian J. Schultheiss, Edward J. Osborne, Vipin T. Sreedharan, André Kahles, Regina Bohnert, Géraldine Jean, Paul Derwent, Paul Kersey, Eric J. Belfield, Nicholas P. Harberd, Eric Kemen, Christopher Toomajian, Paula X. Kover, Richard M. Clark, Gunnar Rätsch, and Richard Mott. Multiple reference genomes and transcriptomes for Arabidopsis thaliana. *Nature*, 477(7365):419–423, sep 2011.

[11] Heng Li and Richard Durbin. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics*, 25(14):1754–1760, jul 2009.

[12] Rayan Chikhi and Guillaume Rizk. Space-efficient and exact de Bruijn graph representation based on a bloom filter. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 7534 LNBI(1):236–248, sep 2012.

[13] G. Rizk, A. Gouin, R. Chikhi, and C. Lemaitre. MindTheGap: integrated detection and assembly of short and long insertions. *Bioinformatics*, 30(24):3451–3457, dec 2014.

[14] Ryan R. Wick, Mark B. Schultz, Justin Zobel, and Kathryn E. Holt. Bandage: Interactive visualization of de novo genome assemblies. *Bioinformatics*, 31(20):3350–3352, oct 2015.

[15] Erwan Drezen, Guillaume Rizk, Rayan Chikhi, Charles Deltel, Claire Lemaitre, Pierre Peterlongo, and Dominique Lavenier. GATB: Genome Assembly & Analysis Tool Box. *Bioinformatics (Oxford, England)*, 30(20):2959–2961, oct 2014.

[16] Cervin Guyomar, Fabrice Legeai, Emmanuelle Jousselin, Christophe Mougel, Claire Lemaitre, and Jean-Christophe Simon. Multi-scale characterization of symbiont diversity in the pea aphid complex through metagenomic approaches. *Microbiome*, 6(1):181, 2018.

[17] Eva Nováková, Václav Hypša, Joanne Klein, Robert G Foottit, Carol D von Dohlen, and Nancy A Moran. Reconstructing the phylogeny of aphids (hemiptera: Aphididae) using dna of the obligate symbiont buchnera aphidicola. *Molecular Phylogenetics and Evolution*, 68(1):42–54, 2013.

[18] Alexey Gurevich, Vladislav Saveliev, Nikolay Vyahhi, and Glenn Tesler. QUAST: Quality assessment tool for genome assemblies. *Bioinformatics*, 29(8):1072–1075, apr 2013.

SPiP: a Splicing Prediction Pipeline addressing the diversity of splice alterations, validated on a curated diagnostic set of 2,784 exonic and intronic variants.

Raphaël LEMAN[1,2,3], Béatrice PARFAIT[4], Dominique VIDAUD[4], Emmanuelle GIRODON[4], Laurence PACOT[4], Gérald LE GAC[5], Chandran KA[5], Claude FEREC[5], Yann FICHOU[5], Céline QUESNELLE[1], Etienne MULLER[1], Dominique VAUR[1], Laurent CASTERA[1], Agathe RICOU[1], Hélène TUBEUF[2], Omar SOUKARIEH[2], Pascaline GAILDRAT[2], Florence RIANT[6], Marine GUILLAUD-BATAILLE[7], Virginie CAUX-MONCOUTIER[8], Nadia BOUTRY-KRYZA[9], Françoise BONNET-DORION[10], Ines SCHULTZ[11], Maria ROSSING[12], Michael T. PARSONS[13], Amanda B. SPURDLE[13], Alexandra MARTINS[2], Claude HOUDAYER[2] , Sophie KRIEGER[1,2,3]

1. Laboratoire de Biologie Clinique et Oncologique, Centre François Baclesse, 14000 Caen, France

2. Inserm U1245 Genomics and Personalized Medecine in Cancer and Neurological Disorders, Normandie Univ, UNIROUEN, Normandy Centre for Genomic and Personalized Medicine, 76031 Rouen, France

3. Normandie Univ, UNICAEN, 14000 Caen, France

4. Service de Génétique et Biologie Moléculaires, APHP, HUPC, Hôpital Cochin, 75014 Paris, France

5. Inserm UMR1078, Genetics, Functional Genomics and Biotechnology, Université de Bretagne Occidentale, 29200 Brest, France

6. Laboratoire de Génétique, AP-HP, GH Saint-Louis-Lariboisière-Fernand Widal, Paris, France

7. Gustave Roussy, Université Paris-Saclay, Département de Biopathologie, 94805 Villejuif, France

8. Service de Génétique, Institut Curie, 75005 Paris, France

9. Unité Mixte de Génétique Constitutionnelle des Cancers Fréquents, Hospices Civils de Lyon, 69000 Lyon, France

10. Institut Bergonie - INSERM U1218 Departement de Biopathologie Unité de Génétique Constitutionnelle, 33000 Bordeaux, France

11. Laboratoire d'Oncogénétique, Centre Paul Strauss, 67000 Strasbourg, France

12. Centre for Genomic Medicine, Rigshospitalet, University of Copenhagen, 1017 Copenhagen, Denmark

13. Department of Genetics and Computational Biology, QIMR Berghofer Medical Research Institute, 4006 Herston, Queensland, Australia

Corresponding Author: S.KRIEGER@baclesse.unicancer.fr

Abstract:

Variant interpretation is recognized as the major challenge in genetic diagnosis. Spliceogenic variants exemplify this issue as all types of nucleotide variations can be pathogenic by affecting normal pre-mRNA splicing via disruption/creation of splicing signals such as splice sites (ss), branchpoints (BPs) or splicing regulatory elements (SREs). Unfortunately, most in silico prediction tools are dedicated to specific signals (eg 5'/3'ss, BPs or SREs). We developed the Splicing Prediction Pipeline (SPiP) to allow comprehensive assessment of variant effect on the different regulatory motifs involved in splicing. SPiP runs a cascade of different and complementary tools, chosen on their efficiency, i.e. Splicing Prediction in Consensus Element

(SPiCE) for physiological 5'/3'ss, Branch Point Prediction (BPP) for BPs, ΔtESRseq for SREs. Moreover, we embedded a new score for the prediction of cryptic/de novo ss, after training and validation on more than 200 million of ss obtained through Ensembl data.

SPiP was evaluated on a curated diagnostic collection of 2,784 variants (13.7% unpublished) in 213 genes, with their corresponding experimental RNA splicing data. These variants were scattered along the exonic and intronic sequences up to 36,947 bp from the ss. SPiP achieved an accuracy of 80.2 %, with a specificity of 70.9 % and a sensitivity of 91.0 %. As a result, SPiP is a comprehensive prediction pipeline which properly deals with the diversity of possible splicing alterations. It can be easily implemented in any diagnostic laboratory as a routine decision making tool for prioritizing RNA studies.

SPiP is available at: https://sourceforge.net/projects/splicing-prediction-pipeline/

Keywords *RNA, variants, splicing predictions, SPiP.*

SPiP: a Splicing Prediction Pipeline addressing the diversity of splice alterations, validated on a curated diagnostic set of 2,784 exonic and intronic variants.

Raphaël LEMAN[1,2,3], Béatrice PARFAIT[4], Dominique VIDAUD[4], Emmanuelle GIRODON[4], Laurence PACOT[4], Gérald LE GAC[5], Chandran KA[5], Claude FEREC[5], Yann FICHOU[5], Céline QUESNELLE[1], Etienne MULLER[1], Dominique VAUR[1], Laurent CASTERA[1], Agathe RICOU[1], Hélène TUBEUF[2], Omar SOUKARIEH[2], Pascaline GAILDRAT[2], Florence RIANT[6], Marine GUILLAUD-BATAILLE[7], Virginie CAUX-MONCOUTIER[8], Nadia BOUTRY-KRYZA[9], Françoise BONNET-DORION[10], Ines SCHULTZ[11], Maria ROSSING[12], Michael T. PARSONS[13], Amanda B. SPURDLE[13], Alexandra MARTINS[2], Claude HOUDAYER[2] , Sophie KRIEGER[1,2,3]

1. Laboratoire de Biologie Clinique et Oncologique, Centre François Baclesse, 14000 Caen, France
2. Inserm U1245 Genomics and Personalized Medecine in Cancer and Neurological Disorders, Normandie Univ, UNIROUEN, Normandy Centre for Genomic and Personalized Medicine, 76031 Rouen, France
3. Normandie Univ, UNICAEN, 14000 Caen, France
4. Service de Génétique et Biologie Moléculaires, APHP, HUPC, Hôpital Cochin, 75014 Paris, France
5. Inserm UMR1078, Genetics, Functional Genomics and Biotechnology, Université de Bretagne Occidentale, 29200 Brest, France
6. Laboratoire de Génétique, AP-HP, GH Saint-Louis-Lariboisière-Fernand Widal, Paris, France
7. Gustave Roussy, Université Paris-Saclay, Département de Biopathologie, 94805 Villejuif, France
8. Service de Génétique, Institut Curie, 75005 Paris, France
9. Unité Mixte de Génétique Constitutionnelle des Cancers Fréquents, Hospices Civils de Lyon, 69000 Lyon, France
10. Institut Bergonie - INSERM U1218 Departement de Biopathologie Unité de Génétique Constitutionnelle, 33000 Bordeaux, France
11. Laboratoire d'Oncogénétique, Centre Paul Strauss, 67000 Strasbourg, France

12.Centre for Genomic Medicine, Rigshospitalet, University of Copenhagen, 1017 Copenhagen, Denmark

13.Department of Genetics and Computational Biology, QIMR Berghofer Medical Research Institute, 4006 Herston, Queensland, Australia

Corresponding Author: S.KRIEGER@baclesse.unicancer.fr

1. Introduction

Since the advent of genome wide sequencing, interpretation of variants of unknown significance (VUS) has been recognized as the major bottleneck and challenge for clinical geneticists. Variants are usually classed within a 5-tiered scheme [1] from benign and likely benign variants (class 1 and 2, respectively) to likely pathogenic and pathogenic variants (class 4 and 5, respectively). The geneticist is on relatively solid ground in these four classes, where the biological impact is known or at least likely known. However, class 3 refers to the so called VUS where the effect of the sequence variation on the transcript and protein and thereby on the patient is simply not known. Clinical management logically stems from this knowledge [2] which is why variant classification is of utmost importance.

Pre-mRNA splicing by the spliceosome is essential for maturation of mRNA. Splicing requires three mandatory motifs on the pre-mRNA molecule, the splice donor site (5'ss), the splice acceptor site (3'ss) and the branch point (BP) (see fig. 1). The 5'ss defines the exon/intron junction at the 5' end of each intron with two highly conserved nucleotides, mainly GT. The consensus motif of 5'ss represents the 3 last nucleotides (nt) of exon and 6 first nt in intron. The 3'ss delineates the intron/exon junction at the 3' end of each intron with a highly conserved dinucleotide (mainly AG). The 12 last nt in intron and 2 first nt in exon constitute the consensus motif of 3'ss [3]. The branch site is a short motif upstream the 3'ss that includes the branch point (BP) adenosine, essential for the splicing process [4]. These BPs are mainly located in area between -44 and -18 nt of the natural 3'ss [5]. Separating the 3'ss and the BPs area, there is a cytosine and thymidine rich sequence called polypyrimidic tract (PPT). The identification of these mandatory motifs depends also of short motifs (6-8 nt) defined as splicing regulatory elements (SREs). Briefly, these motifs are binding signals recognized by RNA-binding proteins, mostly SR (serine and arginine rich) proteins. The SREs can be enhancers or silencers for the identification of splice sites by the spliceosome. The SREs act mostly in exonic region and in this region are called exonic splicing regulatory sequences (ESRseq) [6].

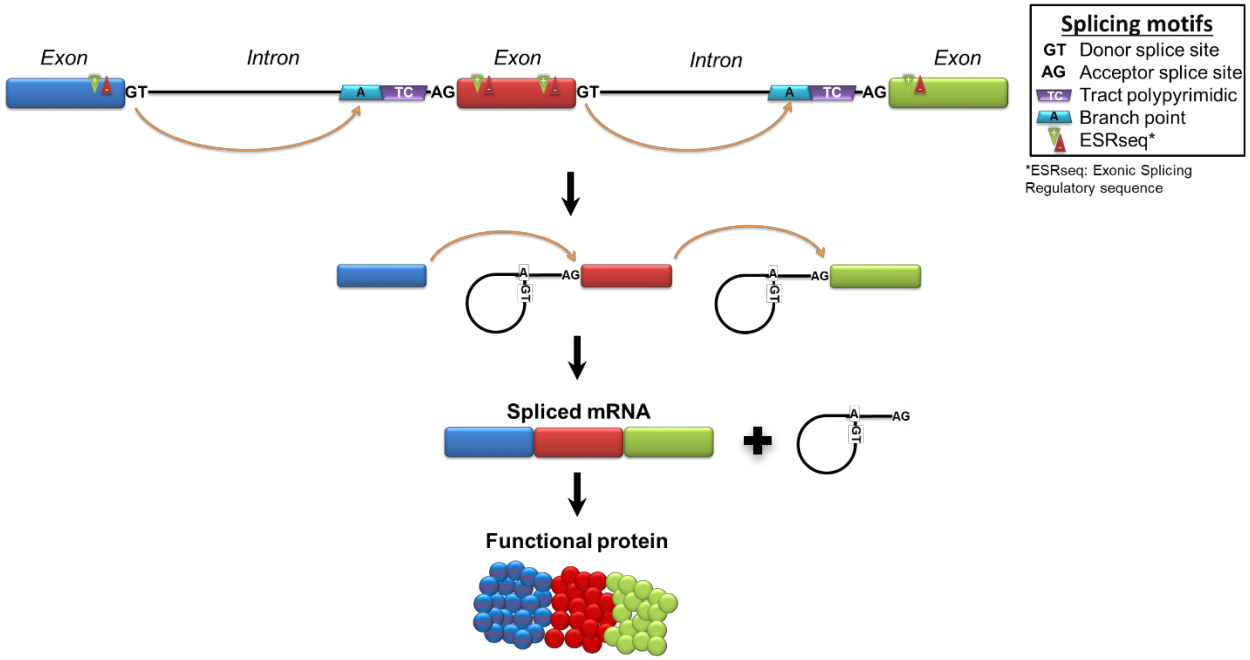


Fig. 1: Splicing mechanism and motifs. A pre-mRNA with three exons, shown as boxes and the introns shown as lines. The donor splice sites are represented by the letters GT, the branch adenosine by A and acceptor splice site by AG. The splicing mechanism ligates exon and released lariat intron to obtain the spliced mRNA.

Spliceogenic variants are probably the most challenging for the geneticists as each nucleotide variation, regardless of its location, can potentially affect pre-mRNA splicing and be pathogenic via disruption/creation of splicing signals such as 5'/3' ss, BPs or SREs. Consequently, assessing the impact of variants on splicing is a mandatory task in molecular diagnosis. Toward this aim, several *in silico* prediction tools were developed. These tools are important to select variants that are worthy of expensive and time-consuming RNA analyses. However, the most bioinformatics tools are dedicated to specific splicing motifs (5'/3' ss, BPs, SREs). We propose in this work a new tool called Splicing Prediction Pipeline (SPiP). SPiP is a decision tree, running a cascade of different and complementary bioinformatics tools (*see below*). Then our tool allows the comprehensive assessment of variant effect on the different regulatory motifs involved in splicing. The main goal of SPiP is to prioritize RNA *in vitro* studies whatever the position of variant is. Moreover, the tool displays the reasons why the variant is predicting as spliceogenic, *i.e.* the splicing motif(s) altered by the variant, to avoid the use of a "black box" prediction tool. Furthermore, and to demonstrate its versatility, SPiP was successfully applied to a set of 2,784 variants, with their corresponding experimental RNA splicing data, occurring in 213 genes.

2. Bioinformatic tools used in SPiP

The bioinformatics tools used by SPiP were chosen on their efficiency to have the optimal tool for each splicing signals (5'/3' ss, BPs, ESRseq). The selection of tool dedicated to 5'/3' ss was performed on the validation set described in Leman *et al.* [7] work (n = 253 variants). These data present the experimental RNA splicing data of these variants occurring in 11 genes. We compared the tools: Splice Site Finder (SSF) [8], MaxEntScan (MES) [9], Human Splicing Finder (HSF) v3.0 [10], ADAboost, RandomForest [11], splicing-based analysis of variants (SPANR) [12] and Splicing Prediction in Consensus Element (SPiCE) [7]. We observed that SPiCE score shown the best performance to predict the alteration of these motifs (see fig. 2). Thus this tool was used by SPiP for the variants in 5'/3' ss.

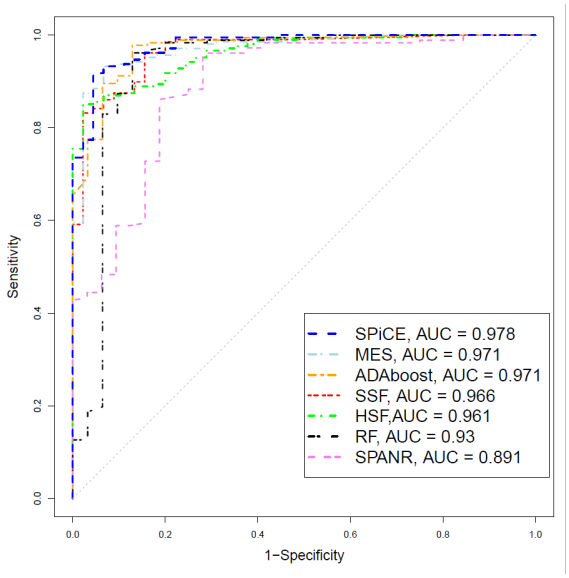


Fig. 2. Comparison of bioinformatics tools 3'/5'ss dedicated on 253 variants.

MES implements the PPT sequence in the score calculation. Indeed, the 3'ss sequence used by MES is located between -20 in intron and +3 in exon. Thus, MES was used by SPiP to study the impact of variant in PPT area.

The optimal tool for branch point prediction was defined on a set of 120 variants with their RNA *in vitro* studies. This data collection was performed in the scope of an in progress benchmarking [13] of 5 BPs-dedicated tools: SVM-BPfinder [14], Branch Point Prediction (BPP) [15], Branchpointer [16], LaBranchoR [17] and RNA Branch Point Selection (RNABPS) [18]. As a result, BPP has shown the best performance (see table 1) to predict a BP alteration. Moreover, the overall best way to consider a BP alteration was to consider a variant as spliceogenic if occurs in the 4-mers motif (TRAY) of the BP predicted by the tools. These 4-mers are the 2 nt upstream the branch A and the nucleotide downstream this A.

Table 1. Classification of variants according their position in the predicted branch point (n = 120) (Motif 4-mers: TRAY). TP (True Positive), FP (False Positive), TN (True Negative), FN (False Negative).

	SVM-BPfinder	BPP	Branchpointer	LaBranchoR	RNABPS
TP	24	32	32	27	30
FP	6	7	12	15	12
TN	76	75	69	67	70
FN	14	6	6	11	8
Accuracy	83.33 %	89.17 %	84.87 %	78.33 %	83.33 %
Sensitivity	63.16 %	84.21 %	84.21 %	71.05 %	78.95 %
Specificity	92.68 %	91.46 %	85.19 %	81.71 %	85.37 %

The optimal tool for the ESRseq, was determined from two independent benchmarking [19,20] performed on a set of experimentally-proven exonic variants. These two studies concluded that ΔtESRseq tool [21] reached the best performance.

For the variant creating new splice site, we developed a new model to detect the use of cryptic splice site. This new model was a metascore gathered the MES, SSF and ESRseq scores [6] for each potential splice site. We trained and validated it on a set splice sites from the transcripts described in Ensembl data (n = 555,679 ss). As control data, we took all AG and GT motifs in these transcripts, corresponding to a comprehensive list of potential splice sites (n = 202,458,596 potential ss). Two third of this data collection was used to train the model, based on logistic regression. One third remaining of this collection was use as validation set. The model reached an area under the ROC curve to 0.974. With the optimal threshold (same values of sensitivity and specificity), the overall accuracy was 92.7 %. The strategy to detect a splice site activation from this model was illustrate in fig. 3. Briefly, we compare the scores of potential splice site around the mutation between wild-type and mutated sequences. From this comparison, the tool considers only the splice sites with a reinforcement of score or a new score apparition to the detection of splice sites. Then on these last splice sites, we applied the optimal threshold previously defined. Whether a splice site has a score above the threshold, we consider it as an activation of new splice site. At the same time of this work, a deep-learning based algorithm, called SpliceRover [22], was published to detect splice site. We compared our new tool with SpliceRover, with our collection of 2,784 variants, with their corresponding experimental RNA splicing data, occurring in 213 genes. As a result, our model has shown the best performances (see table 2).

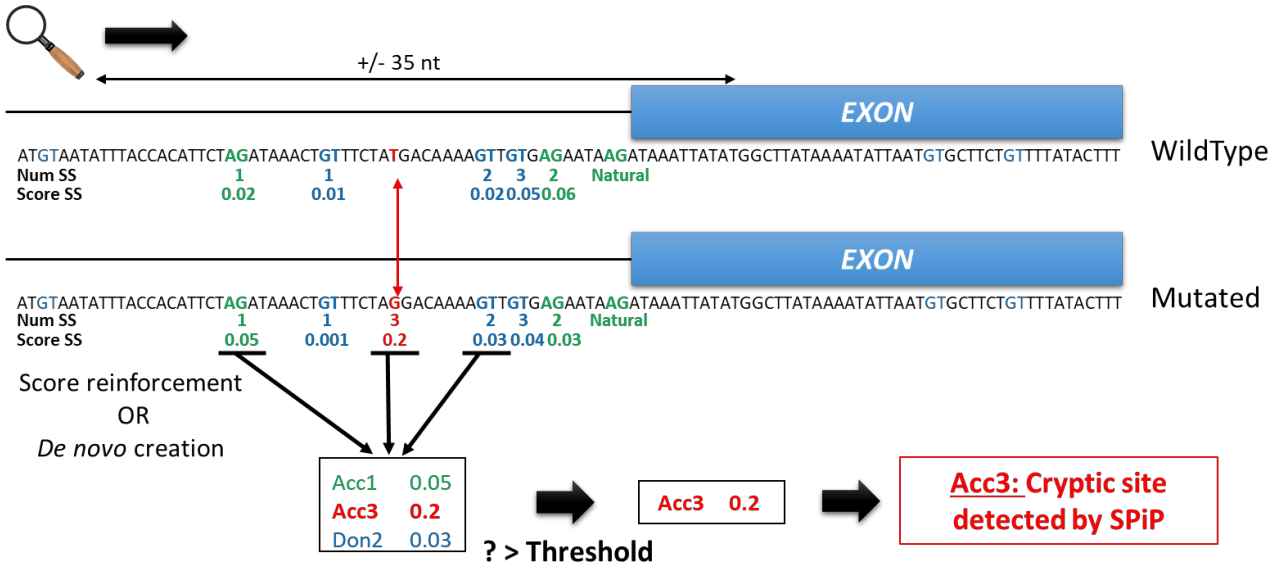


Fig. 3. The strategy used by our tool to detect an activation of cryptic splice site by a mutation. In the scan area the tool detects 2 AG signals and 3 GT signals in wild-type and mutated sequences, plus a third *de novo* AG signal in mutated sequence, (*i.e.* potential splice site). The tool compares the score of each potential

splice sites. Only the second donor site and the first acceptor sites have a reinforcement of score plus the *de novo* acceptor site (Acc3). On these 3 splice sites, only the *de novo* splice site has a score above the decisional threshold (see text) and so is predicted as cryptic splice site.

Table 2. Comparison between SpliceRover [22] and our model to the detection of cryptic splice site activation, N = 2,784 variants. AUC_{ROC}: Area Under the ROC Curves

	SpliceRover	**Our model**
AUC_{ROC}	82.8 %	86.8 %
Accuracy	75.2 %	84.6%
Sensitivity	74.5 %	78.1 %
Specificity	75.2 %	85.3 %

3. SPiP workflow

For each variant processed by SPiP, the tool determines, firstly, this position in the transcript to select the relevant tool (see fig. 4). Secondly, SPiP apply the optimal tools for these splicing motifs to predict or not a splicing alteration. With an exception for the new splice site activation, SPiP checks the apparition of this event whatever the position of the variant. The tool ΔtESRseq was not use by SPiP if the variant occurred at more than 120 nt of splice site in exon. Indeed, this tool uses the ESRseq scores obtained from minigene assays and the ESRseqs located at more than 120 nt were not analyzed. To illustrate the SPiP running, we can take the variant located in c.2410-18C>G in *NF1* gene (NM_000267). In this example, the variant was located in the PPT area and in the BP area. Therefore, SPiP used MES, BPP and researched a new splice site activation. For this variant, MES predicts an alteration of the 3'ss. The BP predicted by BPP was located in -43 of the natural 3'ss, so the variant was outside the motif of this BP. This variant creates also a new potential 3'ss by making an AG signal, instead of AC in the wild-type sequence. This potential 3'ss was detected as cryptic splice site by our model. Then the overall prediction of SPiP was an alteration of the PPT motif and a creation of new 3'ss. The true splicing effect of this variant observed by RNA *in vitro* study from peripheral blood RNA sample was the retention of the 17 last nt of the intron by activation of a new 3'ss [23].

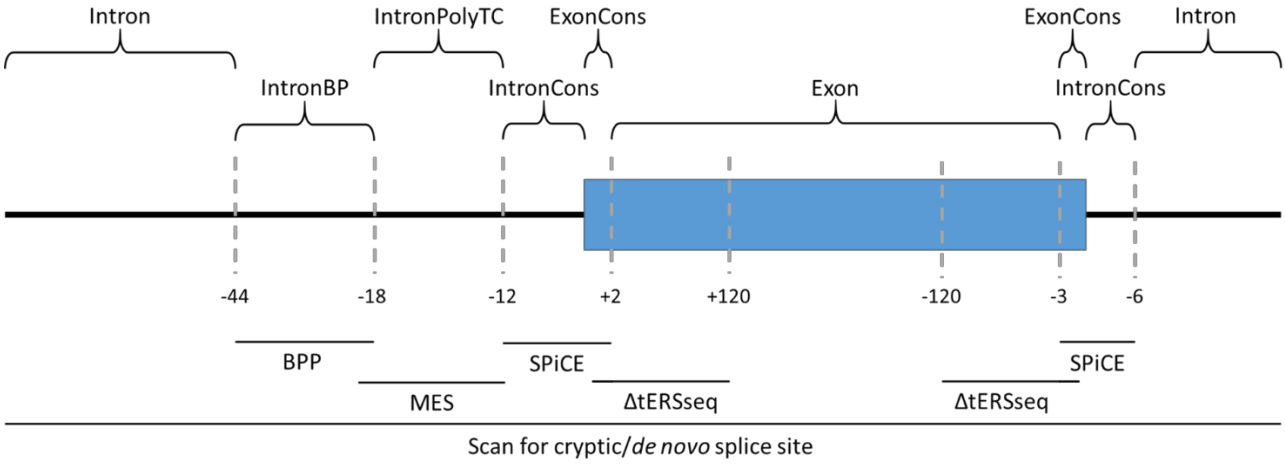


Fig. 4. The bioinformatics tools used for each splicing motifs. IntronBP: Branch point area, IntronPolyTC: polypyrimidine tract, IntronCons: Intronic consensus splice site, ExonCons: Exonic consensus splice site.

4. SPiP running

To ensure an access of SPiP, we developed an R script to calculate the scores (available at: https://sourceforge.net/projects/splicing-prediction-pipeline/). Thus installation of other bioinformatics tools is not necessary to calculate scores. From the mutation name, this script permits to determine the position of the variant in the transcript, to calculate the relevant scores. Then SPiP generates a synthesis of potential impact of variant at splicing level. The transcripts database used by SPiP is the RefSeq database with the assembly

genome version hg19 and hg38. The input of SPiP is the transcript ID (RefSeq) and the HGVS (Human Genome Variation Society) mutation nomenclature, ":"-separated, (ex: NM_007294:c.4096+3A>G). SPiP can also deal with the Variant Call Format (VCF) v4.0 or later, a standardized text file format for representing mutation. We propose two versions of this tool, Windows version and Linux version. The Windows version is an interface software available in standalone version, thus not supplemental installation is necessary to use SPiP. The Linux version was developed to treat a great number of variant. Indeed, the analysis can be parallelizable on several threads to reduce the time of calculation. This version needs a R environment (v3.0 or later) with the librairies 'Rcurl' and 'parallel' and the tool samtools v1.6 or later [24]. The runtime of Windows version was 3 variants by second on AMD Ryzen 7 PRO 1700 Eight-Core processor. The runtime of Linux version was 0.415 second by variant and by thread on Intel(R) Xeon(R) CPU E5-2670 v2 @ 2.50GHz.

5. Evaluation of SPiP with 2,784 variants

To evaluate SPiP, we collected 2,784 variants with their RNA *in vitro* studies. These variants occurred in 213 genes involved in human genetic disorders. The data were from the literature (n = 2,403 variants) and the 381 remaining variants were not published data. These last variants were studied for diagnosis purpose and collected by a collaborative effort of the French splicing network of Unicancer Genetic Group (UGG), the institute Cochin, the Inserm U1078, the Inserm U1245, the laboratory of genetic of the hospital Saint-Louis-Lariboisière-Fernand Widal and the Center of genomics of the university of Copenhagen. Among these 2,784 variants 53 % (1,490/2,784) had not impact on splicing. The 1,294 spliceogenic variants induced exon skipping (64.2 %; 831/1,294), the use of new splice site (27.7 %; 359/1,294), the exonisation of intron (6.8 %; 88/1,294) and the total intron retention (1.2 %; 16/1,294). The reparation of variant in the different area of transcript was illustrated in fig. 5. The most distant variant of the splice site was the mutation c. 31+36947G>A in *DMD* gene (NM_004006), located at 36,947 nt of the natural 5'ss in intron.

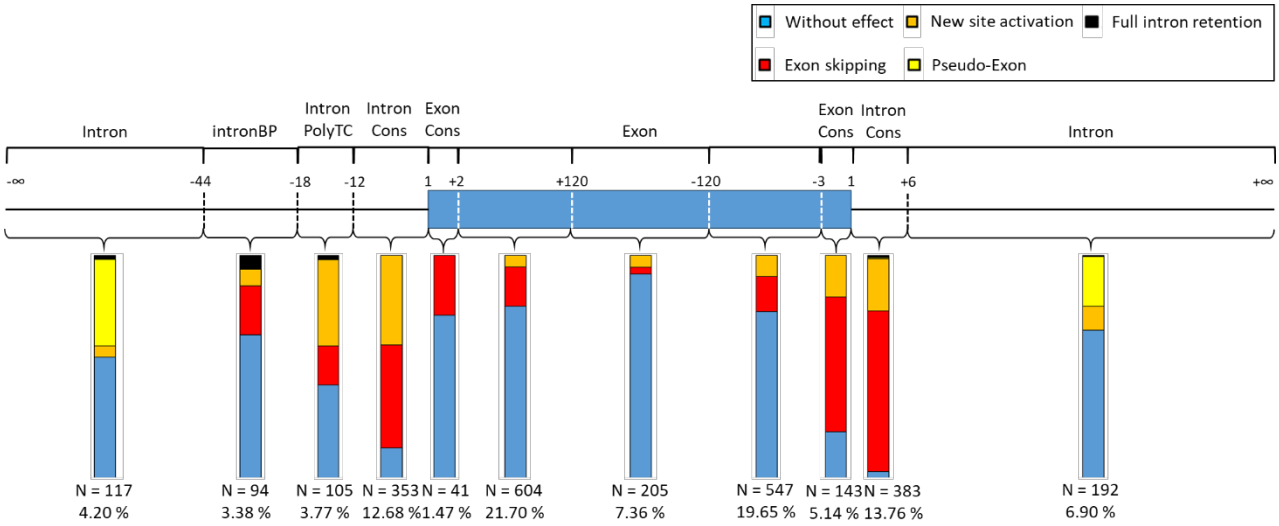


Fig. 5. Repartition of variants in the different splicing motif, N = 2,784 variants.

The overall accuracy of SPiP was 80.21 %, the sensitivity was 90.96 % and the specificity was 70.87 %. The SPiP sensitivities to detect an exon skipping, the use of new splice site, the exonisation of intron and the intron retention were 89.77 %; 96.03 %; 82.96 %; 81.25 %, respectively. These performances highlight the capability of SPiP to detect an alteration of splicing for each splicing motif, whatever the position of variant and independently of the gene. Therefore, SPiP has the potential of a widely used decision-making tool to guide geneticists toward relevant spliceogenic variants in the deluge of high-throughput sequencing data.

References

1. Plon SE, Eccles DM, Easton D, Foulkes WD, Genuardi M, Greenblatt MS, et al. Sequence variant classification and reporting: recommendations for improving the interpretation of cancer susceptibility genetic test results. Human Mutation. 2008;29(11):1282–91.

2. Richards S, Aziz N, Bale S, Bick D, Das S, Gastier-Foster J, et al. Standards and Guidelines for the Interpretation of Sequence Variants: A Joint Consensus Recommendation of the American College of Medical Genetics and Genomics and the Association for Molecular Pathology. Genet Med. 2015 May;17(5):405–24.

3. Burge CB, Tuschi T, Sharp PA. Splicing of Precursors to mRNAs by the Spliceosomes. In: The RNA World II. Cold Spring Harbor Laboratory Press; 1999. p. 525–60.

4. Will CL, Lührmann R. Spliceosome Structure and Function. Cold Spring Harb Perspect Biol. 2011 Jan 7;3(7):a003707.

5. Mercer TR, Clark MB, Andersen SB, Brunck ME, Haerty W, Crawford J, et al. Genome-wide discovery of human splicing branchpoints. Genome Res. 2015 Jan 5;gr.182899.114.

6. Ke S, Shang S, Kalachikov SM, Morozova I, Yu L, Russo JJ, et al. Quantitative evaluation of all hexamers as exonic splicing elements. Genome Res. 2011 Jan 8;21(8):1360–74.

7. Leman R, Gaildrat P, Gac GL, Ka C, Fichou Y, Audrezet M-P, et al. Novel diagnostic tool for prediction of variant spliceogenicity derived from a set of 395 combined in silico/in vitro studies: an international collaborative effort. Nucleic Acids Res. 2018 Sep 6;46(15):7913–23.

8. Shapiro MB, Senapathy P. RNA splice junctions of different classes of eukaryotes: sequence statistics and functional implications in gene expression. Nucleic Acids Res. 1987 Sep 11;15(17):7155–74.

9. Yeo G, Burge CB. Maximum Entropy Modeling of Short Sequence Motifs with Applications to RNA Splicing Signals. Journal of Computational Biology. 2004 Mar 1;11(2–3):377–94.

10. Desmet F-O, Hamroun D, Lalande M, Collod-Béroud G, Claustres M, Béroud C. Human Splicing Finder: an online bioinformatics tool to predict splicing signals. Nucleic Acids Res. 2009 May 1;37(9):e67–e67.

11. Jian X, Boerwinkle E, Liu X. In silico prediction of splice-altering single nucleotide variants in the human genome. Nucleic Acids Res. 2014 Dec 16;42(22):13534–44.

12. Xiong HY, Alipanahi B, Lee LJ, Bretschneider H, Merico D, Yuen RKC, et al. The human splicing code reveals new insights into the genetic determinants of disease. Science. 2015 Jan 9;347(6218):1254806.

13. Leman R, Tubeuf H, Raad S, Tournier I, Derambure C, Lanos R, et al. Assessment of branch point bioinformatics tools to predict physiological branch points and their alteration by variants. Article in progress

14. Corvelo A, Hallegger M, Smith CWJ, Eyras E. Genome-Wide Association between Branch Point Properties and Alternative Splicing. PLOS Computational Biology. 2010 Nov 24;6(11):e1001016.

15. Zhang Q, Fan X, Wang Y, Sun M, Shao J, Guo D, et al. BPP: a sequence-based algorithm for branch point prediction. Bioinformatics. 2017 Oct 15;33(20):3166–72.

16. Signal B, Gloss BS, Dinger ME, Mercer TR, Hancock J. Machine learning annotation of human branchpoints. Bioinformatics. 2018 Mar 15;34(6):920–7.

17. Paggi JM, Bejerano G. A sequence-based, deep learning model accurately predicts RNA splicing branchpoints. RNA. 2018 Sep 17;rna.066290.118.

18. Nazari I, Tayara H, Chong KT. Branch Point Selection in RNA Splicing Using Deep Learning. IEEE Access. 2019;7:1800–7.

19. Soukarieh O, Gaildrat P, Hamieh M, Drouet A, Baert-Desurmont S, Frébourg T, et al. Exonic Splicing Mutations Are More Prevalent than Currently Estimated and Can Be Predicted by Using In Silico Tools. PLOS Genetics. 2016 Jan 13;12(1):e1005756.

20. Grodecká L, Buratti E, Freiberger T. Mutations of Pre-mRNA Splicing Regulatory Elements: Are Predictions Moving Forward to Clinical Diagnostics? International Journal of Molecular Sciences. 2017 Aug;18(8):1668.

21. Giacomo DD, Gaildrat P, Abuli A, Abdat J, Frébourg T, Tosi M, et al. Functional Analysis of a Large set of BRCA2 exon 7 Variants Highlights the Predictive Value of Hexamer Scores in Detecting Alterations of Exonic Splicing Regulatory Elements. Human Mutation. 2013;34(11):1547–57.

22. Zuallaert J, Godin F, Kim M, Soete A, Saeys Y, De Neve W. SpliceRover: interpretable convolutional neural networks for improved splice site prediction. Bioinformatics. 2018 Dec 15;34(24):4180–8.

23. Xu W, Yang X, Hu X, Li S. Fifty-four novel mutations in the NF1 gene and integrated analyses of the mutations that modulate splicing. International Journal of Molecular Medicine. 2014 Jul 1;34(1):53–60.

24. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, et al. The Sequence Alignment/Map format and SAMtools. Bioinformatics. 2009 Aug 15;25(16):2078–9.

Architecture and evolution of blade assembly in β-propeller lectins

François Bonnardel [1,2,3], Atul Kumar [1,4], Michaela Wimmerova [4,5], Martina Lahmann [6], Serge Perez [1], Annabelle Varrot [1], Frédérique Lisacek [2,3*], and Anne Imberty [1*]

[1] Univ. Grenoble Alpes, CNRS, CERMAV, 38000 Grenoble, France.

[2] Swiss Institute of Bioinformatics, CH-1227 Geneva, Switzerland.

[3] Computer Science Department, UniGe, CH-1227 Geneva, Switzerland.

[4] CEITEC, Masaryk University, 625 00 Brno, Czech Republic

[5] NCBR, Fac.Sci, Masaryk University, 625 00 Brno, Czech Republic

[6] School of Chemistry, University of Bangor, LL57 2UW Bangor, United Kingdom

Corresponding Author: francois.bonnardel@cermav.cnrs.fr

Paper Reference: **Bonnardel *et al.* (2019) Architecture and Evolution of Blade Assembly in β-propeller Lectins. Structure, Cell press. In press, 2019.** https://doi.org/10.1016/j.str.2019.02.002

Abstract: *Lectins with a β-propeller fold bind glycans on the cell surface through multivalent binding sites and appropriate directionality. These proteins are formed by repeats of short domains, raising questions about evolutionary duplication. However, these repeats are difficult to detect in translated genomes and seldom correctly annotated in sequence databases. To address these issues, we defined the blade signature of the five types of β-propellers using 3D-structural data. With these templates, we predicted 3887 β-propeller lectins in 1889 species and organised this new information in a searchable online database. The data reveals a widespread distribution of β-propeller lectins across species. Prediction also emphasises multiple architectures and led to uncover a novel β-propeller assembly scenario. This was confirmed by producing and characterizing a predicted protein coded in the genome of Kordia zhangzhouensis. The crystal structure shows a new intermediate in the evolution of β-propeller assembly, demonstrates the power of designing bioinformatics tools and the benefit of multidisciplinary interactions.*

Keywords: β-propeller, lectin, oligomerisation, carbohydrate binding protein, prediction, motif

1. Introduction

Lectins are protein receptors that can bind at least one carbohydrate, and with no enzymatic function [1]. Lectins are generally multivalent and such multiplicity of carbohydrate binding sites favours the strong avidity to glycoconjugates available in multiple copies on all cell surfaces. Lectins are involved in a range of biological processes taking place between cells. For example, they participate in the interaction between microorganisms and hosts cells (pathogenicity, symbiosis…). Despite such a prevalent role, lectins are rather poorly characterised in protein databases. To overcome this shortcoming, we launched the Unilectin3D database [2] that includes a large number of classified and manually curated lectin 3D-structures, with information on their fold, oligomeric structure and carbohydrate binding site(s). The Unilectin3D collection highlights the diversity of folds that lectins adopt, and the high frequency of the occurrence multimeric structures. However, for some lectins, multivalency is not created by oligomerization, but by tandem repeat of conserved carbohydrate binding domains. Such tandem repeats are observed in the so-called β-propeller lectins. The β-propeller is a fold widely distributed in Nature. β-propeller proteins adopt a donut shape made of four to ten repeats (or blades) of four-stranded β-sheets [3]. Their functions are broad, generally related to an enzymatic active site located in the centre of the structure. Although very variable in amino acid sequences, β-propellers have been proposed to derive from a single peptide through multiple episodes or duplication and diversification. The β-propellers proteins (PropLec) of Unilectin3D have been classified in seven different groups. CATH-GENE3D has categories for propellers from 3 to 8 blades, yet not all PropLecs are included. In fact, β-propeller lectins are difficult to identify based on their amino acid sequence. The presence of short repeated peptide motifs (30 to 50 amino acids) challenges classical search programs that are based on sequence alignment. This setback in turn, impacts the definition of protein family in Pfam which defines profiles based on domain similarity. Pfam

profiles matching PropLecs cover either part(s) of one blade (46-58 amino acids) or the whole propeller. As a result, no current tool can, as is, efficiently mine β-propellers, and they usually miss the conserved carbohydrate binding sites of PropLecs.

2. Identification of b-propeller lectin families and prediction

The presence of repeated domains in PropLecs challenges their automatic detection in genomes. Our strategy was to turn this into an advantage by defining conserved motifs corresponding to the blade signature in each family, and then to search multiple and successive occurrences of these motifs in genomes. The seven sub-groups of PropLecs that are described in Unilectin3D were defined based on structural similarity and taxonomy. By focusing only on structural and sequence similarity, we reduced this number to five PropLec families. To simplify the nomenclature, each family has been named according to the number of constituting blades, e.g. PropLec5A, PropLec6A, PropLec6B, PropLec7A and PropLec7B. The structural information in the 13 different PropLecs that have been crystallized so far was used to identify the blade signature of each PropLec family. The peptide sequences were first processed with the RADAR software [4] in order to align the repeated regions. This alignment was refined on the basis of 3D-structural information, which entailed the adjustment of repeat boundaries to the definition of blades. When necessary, alignments were shifted along the sequence so as to centre each blade on the 3D structure. The resulting blade sequence alignments served as the basis for determining conserved motifs and defining characteristic profiles in the form of Hidden Markov Models (HMM). These models were generated with the HMMbuild [5]. HMM profiles identify similar domains depending on the amino acid frequencies at each position of the blade and on the amino acids in previous positions. In order to identify PropLecs in other organisms, the designed motifs were fed into HMMSEARCH to process the UniRef100 non-redundant protein database. The predicted protein sequences were filtered with an e-value set to 0.01 while other parameters were left to default values. This search returned 3877 putative PropLec sequences. A dedicated interface for mining the PropLec database is available at https://www.unilectin.eu/propeller/.

3. Exploration of predicted propeller and experimental validation

β-propellers are generally consisting of one peptide presenting a tandem-repeat. The only exception occurred in the PropLec6A family: these lectins have been characterized in three fungi with six blade repeats for a domain approximately 300 amino acid-long, but also in bacteria with two blade repeats in a 90 amino acid domain, that trimerizes to form the same 6-blade propeller. This is the only case of natural β-propeller assembled by oligomerization. The bimodal distribution of blade numbers in PropLec6A family, with maxima at 6-blade and 2-blade is shown in Figure 5 and in supplemental information. However, from the graph distribution, we predicted that 3-blade domains could also exist, which would correspond to a β-propeller formation by dimerization that was never observed before. The predicted 3-blade sequences of PropLec6A were therefore analysed to select those with a high similarity score, an approximate size of 150 amino acids (three repeats) and correct gene start and ending. Four sequences were selected and annotated as 3-blades lectins from *Kordia zhangzhouensis*, *K. periserrulae*, *Penicillium polonicum* and *P. freii*. The b-propeller protein from the gene KozL has then been expressed, purified and crystallized, allowing to analyze and obtain the 3D structure of a 3-blade protein assembling in dimer to form a complete propeller, and in tetramer with two superposed propellers.

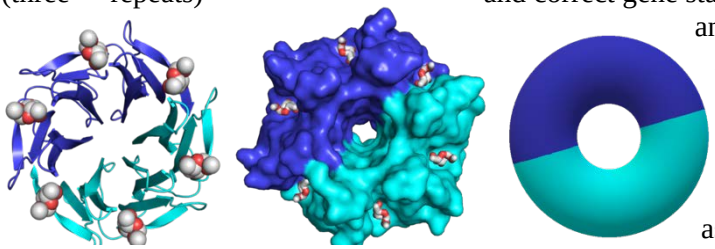


Figure 1: experimentally validated β-propeller dimer protein

References

[1] H. Lis and N. Sharon, Lectins: Carbohydrate-Specific Proteins That Mediate Cellular Recognition. *Chem. Rev.* 1998.

[2] F. Bonnardel et al. UniLectin3D, a database of carbohydrate binding proteins with curated information on 3D structures and interacting ligands. *Nucleic Acids Res.* 2019.

[3] C. Chen et al. The many blades of the β-propeller proteins: Conserved but versatile. *Trends Biochem. Sci.* 2011.

[4] A. Heger and L. Holm. Rapid automatic detection and alignment of repeats in protein sequences. *Proteins Struct. Funct. Genet.* 2000.

[5] R. D. Finn et al. HMMER web server: 2015 Update. *Nucleic Acids Res.* 2015.

Deciphering intra-species bacterial diversity of meat and seafood spoilage microbiota using *gyrB* amplicon sequencing: A comparative analysis with *16S* rDNA V3-V4 amplicon sequencing

Simon POIRIER[1], Olivier RUÉ[2], Raphaëlle PEGUILHAN[1], Gwendoline Coeuret[1], MONIQUE ZAGOREC[3], MARIE-CHRISTINE CHAMPOMIER-VERGÈS[1], VALENTIN LOUX[2], STÉPHANE CHAILLOU[1]

[1] MICALIS, INRA, AgroParisTech, Université Paris-Saclay, Jouy-en-Josas, France
[2] MaIAGE, INRA, Université Paris-Saclay, Jouy-en-Josas, France
3 Secalim, INRA, Oniris, Nantes, France

Corresponding Author: olivier.rue@inra.fr

Paper Reference: **Poirier S, Rué O, Peguilhan R, Coeuret G, Zagorec M, Champomier-Vergès M-C, et al. (2018) Deciphering intra-species bacterial diversity of meat and seafood spoilage microbiota using *gyrB* amplicon sequencing: A comparative analysis with *16S* rDNA V3-V4 amplicon sequencing. PLoS ONE 13(9): e0204629. https://doi.org/10.1371/journal.pone.0204629**

Abstract *Meat and seafood spoilage ecosystems harbor extensive bacterial genomic diversity that is mainly found within a small number of species but within a large number of strains with different spoilage metabolic potential. To decipher the intraspecies diversity of such microbiota, traditional metagenetic analysis using the* 16S *rRNA gene is inadequate. We therefore assessed the potential benefit of an alternative genetic marker,* gyrB*, which encodes the subunit B of DNA gyrase, a type II DNA topoisomerase. A comparison between* 16S *rDNA-based (V3-V4) amplicon sequencing and* gyrB*-based amplicon sequencing was carried out in five types of meat and seafood products, with five mock communities serving as quality controls. Our results revealed that bacterial richness in these mock communities and food samples was estimated with higher accuracy using* gyrB *than using* 16S *rDNA. However, for Firmicutes species, 35% of putative* gyrB *reads were actually identified as sequences of a* gyrB *paralog,* parE*, which encodes subunit B of topoisomerase IV; we therefore constructed a reference database of published sequences of both* gyrB *and pare for use in all subsequent analyses. Despite this co-amplification, the deviation between relative sequencing quantification and absolute qPCR quantification was comparable to that observed for* 16S *rDNA for all the tested species. This confirms that* gyrB *can be used successfully alongside* 16S *rDNA to determine the species composition (richness and evenness) of food microbiota. The major benefit of* gyrB *sequencing is its potential for improving taxonomic assignment and for further investigating OTU richness at the subspecies level, thus allowing more accurate discrimination of samples. Indeed, 80% of the reads of the* 16S *rDNA dataset were represented by thirteen* 16S *rDNA-based OTUs that could not be assigned at the species-level. Instead, these same clades corresponded to 44* gyrB*-based OTUs, which differentiated various lineages down to the subspecies level. The increased ability of* gyrB*-based analyses to track and trace phylogenetically different groups of strains will generate improved resolution and more reliable results for studies of the strains implicated in food processes.*

Keywords Metabarcoding, gyrase subunit B, subspecies level, meat and seafood spoilage.

Despite the extraordinary insights that have been gained through *16S* rDNA profiling analyses, taxonomic methods based on this approach have several shortcomings, particularly at the shallowest taxonomic levels. The extremely slow rate of evolution of this gene hinders the resolution of closely related bacteria into individual *16S* rDNA phylotypes. Moreover, variation in the number of rRNA operons among different bacterial species creates problems for the quantification of cell numbers or taxon abundances based on *16S* rDNA phylotypes [1]. For these reasons, *16S* rDNA amplicon data are often analyzed at the genus level only, but these results lack the power to yield informative answers to many questions. Because of this, we sought

an alternative marker that could improve diversity analysis at the species- or even intraspecies-level while keeping the ease-of-use and cost-effectiveness of amplicon sequencing.

Among these, *gyrB* has a higher rate of base substitution than *16S* rDNA does, and shows promise for community-profiling applications [2]. This gene is essential and ubiquitous in bacteria and is sufficiently large in size for use in analysis of microbial communities. It is a single-copy housekeeping gene that encodes the subunit B of DNA gyrase, a type II DNA topoisomerase, and therefore plays an essential role in DNA replication. Furthermore, the *gyrB* gene is also present in Eukarya and sometimes in Archaea but it shows enough sequence dissimilarity between the three domains of life to be used selectively for Bacteria [3].

The main objective of the current work was to validate the usefulness of *gyrB* as an alternative phylogenetic marker to accurately and precisely discriminate closely related species within various food microbiota. We therefore carried out a comparison of amplicon sequencing based on *16S* rDNA V3-V4 and that based on *gyrB* using five types of meat and seafood products (pork sausage, poultry sausage, cod filet, salmon filet, and ground beef). These products were specifically chosen because their microbiota have been extensively studied [4] and comprise a broad spectrum of bacterial species from the phyla Firmicutes and Proteobacteria. In order to assess the added value brought by *gyrB* sequencing with respect to *16S* rDNA sequencing, five mock communities (MC) were constructed as quality controls, using 15 different species with a high degree of intraspecies diversity.

Our results demonstrate that *gyrB* sequencing can fulfill this goal. This housekeeping gene shows around 94 to 95% sequence identity among strains of the same species, a level of variation that matches the ANI (Average Nucleotide Index) value now commonly used for species-level estimation. This ability to distinguish among groups of phylogenetically distinct strains (population lineages, main clonal complexes, and so forth) has enormous implications for our knowledge of the bacterial strains and population fluctuations involved in food processes.

A second objective was to validate a methodological approach to build a *gyrB* databank from public and private resources without introducing ambiguity and redundancy by giving preferences to some curated databases compared to some general but diverse databases.

Our opinion is that *gyrB* sequencing would be very valuable in analyses of bacterial diversity that are specifically directed at deciphering details of population structure at the subspecies level. This approach would carry notable benefits for the temporally and/or spatially extensive campaigns that are often carried out on food microbiota, e.g., studies that track and trace whether particular subspecies lineages are specifically selected or subjected to seasonal changes within a food production chain or during the shelf life. Nevertheless, a knowledge of *gyrB* subspecies lineages sequences is needed to obtain comparable results. The construction of the databank from annotated genomes is thus a key-step.

However, we believe that *16S* rDNA amplicon sequencing should still be incorporated in these metagenetic analyses as a control (by selecting a subset of samples for instance) in order to ensure that the *gyrB* data remain consistent with those of the universally used *16S* rDNA. Therefore, we would not recommend the use of *gyrB*-based methods to de novo analyze microbiota that are completely unknown. Generally speaking, *gyrB* sequencing still needs to be tested in many different types of complex microbiota and especially in those that contain phyla other than Firmicutes and Proteobacteria.

Acknowledgements

This work is supported by the ANR project REDLOSSES (http://www.redlosses.fr).

References

[1] Sun DL, Jiang X, Wu QL, Zhou NY. Intragenomic heterogeneity of *16S* rRNA genes causes overestimation of prokaryotic diversity. Applied and environmental microbiology. 2013;79(19):5962–9. pmid:23872556.

[1] Watanabe K, Nelson J, Harayama S, Kasai H. ICB database: the *gyrB* database for identification and classification of bacteria. Nucleic Acids Research. 2001;29(1):344–5. pmid:11125132.

[2] Forterre P, Gadelle D. Phylogenomics of DNA topoisomerases: their origin and putative roles in the emergence of modern organisms. Nucleic Acids Res. 2009;37(3):679–92. pmid:19208647.

[3] Chaillou S, Chaulot-Talmon A, Caekebeke H, Cardinal M, Christieans S, Denis C, et al. Origin and ecological selection of core and food-specific bacterial communities associated with meat and seafood spoilage. The ISME journal. 2015;9(5):1105–18. pmid:25333463.

[elPrep 4: A high-performance tool for sequence analysis]

Charlotte Herzeel[1] and Pascal Costanza[1]
Imec, ExaScience Lab, Kapeldreef 75, 3001, Leuven, Belgium

Corresponding author: charlotte.herzeel@imec.be

Reference paper: **Herzeel *et al.* (2019) elPrep 4: A multithreaded framework for sequence analysis. *PLOS One*.** https://doi.org/10.1371/journal.pone.0209523

Abstract *We present an overview of elPrep, a framework for processing sequence alignment/map files in the Go programming language, which is developed as a drop-in replacement for GATK, Picard, and SAMtools functionality. Our latest release, elPrep 4, includes multiple features allowing the user to process the preparation steps defined by the GATK Best Practice pipelines for variant calling. This includes new and improved functionality for sorting, (optical) duplicate marking, base quality score recalibration, BED and VCF parsing, and various filtering options. The implementations of these options in elPrep 4 faithfully reproduce the outcomes of their counterparts in GATK 4, SAMtools, and Picard, even though the underlying algorithms are redesigned to take advantage of elPrep's parallel execution framework to vastly improve the runtime and resource use compared to these tools. Our benchmarks show that elPrep executes the preparation steps of the GATK Best Practices up to 13x faster on WES data, and up to 7.4x faster for WGS data compared to running the same pipeline with GATK 4, while utilizing fewer compute resources.*

Keywords NGS, software pipelines, SAM/BAM, parallel programming.

1 Overview

elPrep 4 [1] is a reimplementation of the elPrep framework [2] for processing sequence alignment/map files (SAM/BAM) in the Go programming language [3]. elPrep 4 includes all of the tools that are necessary for implementing the GATK Best Practices pipelines for variant calling, which typically consist of sorting, PCR/optical duplicate marking, and base quality score recalibration and application. A working elPrep command for running such a pipeline is shown in Listing 1. The elPrep tools for the different operations produce outputs that are identical to those of the original tools in GATK, Picard, and SAMTools, while greatly speeding up the execution time.

```
elprep sfm input.bam output.bam
 --mark-duplicates --mark-optical-duplicates output.metrics
 --sorting-order coordinate
 --bqsr output.recal
 --known-sites dbsnp_138.hg38.elsites
 --bqsr-reference hg38.elfasta
```

Listing 1. elPrep command for executing a GATK Best Practices preparation pipeline.

2 Implementation

elPrep is developed at the ExaScience Lab at imec and released as an open-source project on GitHub at https://github.com/ExaScience/elprep under the GNU Affero General Public License version 3 as published by the Free Software Foundation, with Additional Terms.

3 Benchmarks

We use a 4-step pipeline from the GATK Best Practices for benchmarking the computational efficiency of elPrep. Concretely, we set up experiments where we compare the raw performance of elPrep 4 to GATK 4 and GATK 3.8, as well as a scaling experiment on Amazon Web Services to also compare the dollar cost of running elPrep 4 versus GATK 4. We report benchmarks for both a public whole-exome and whole-genome sequencing of NA12878.

The results for whole-exome data are shown in Fig. 1. We show three graphs, comparing the runtime and the RAM and disk use for GATK 4 and elPrep 4 respectively. The runtime graph shows the runtime for each individual pipeline step in the case of GATK 4 (top) versus a combined runtime for elPrep 4 (bottom), as elPrep merges the execution of the different pipeline steps. elPrep allows two execution modes, either running entirely in RAM (filter), or splitting and processing the data by chromosomal regions (sfm), hence there are two results for elPrep. There is no difference in terms of output between either modes. The runtime graph shows elPrep executes the pipeline between 5.4-13x faster. The second graph in Fig. 1 shows the peak RAM use is between 0.7-2.6x of the RAM use in GATK 4, depending on which elPrep mode is used. Similarly, elPrep 4 uses between 0.2-0.6x of the peak disk use that GATK 4 uses. We present similar benchmark graphs for 50x whole-genome data in Fig. 2. Overall, elPrep 4 executes the 4-step pipeline in 3h27h versus 27h in GATK 4, using 192 GB RAM versus 229 GB in GATK 4, and 364 GB peak disk space versus 520 GB in GATK 4.

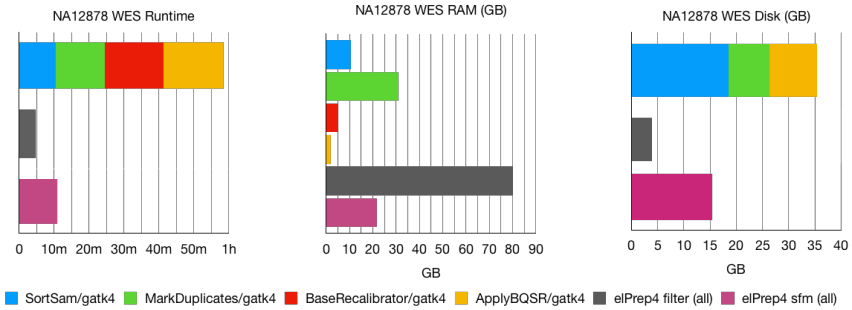


Fig. 1. WES benchmarks. Runtime, RAM use, and disk use in GATK 4 vs. elPrep 4 (filter mode) vs. elPrep 4 (sfm mode). We see 5.4-13x speedup for 0.7-2.6x RAM use and 0.6-0.2x disk use when comparing elPrep 4 filter/sfm to GATK 4. The results, i.e. final BAM, metrics and recalibration files, are the same for all runs.

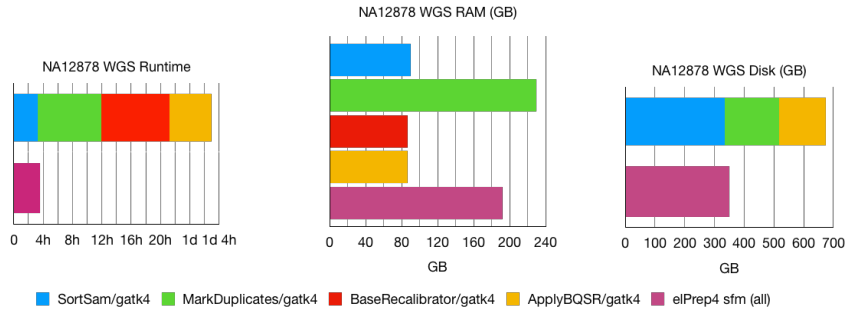


Fig. 2. WGS benchmarks. Runtime, RAM use, and disk use in GATK 4 vs. elPrep 4 (sfm mode). elPrep 4 executes the pipeline 7.4x faster than GATK 4, using 0.84x of the RAM, and only 0.7x of the disk space. The final BAM, metrics, and recalibration files are the same for both runs.

We also report benchmarks to compare elPrep 4 to GATK 3.8 (not shown here). elPrep 4 executes the 4-step pipeline 18.2x faster than GATK 3.8, using only 0.85x of the RAM and 0.8x of the disk space that GATK 3.8 uses. The outputs from elPrep 4 and GATK 3.8 are not identical because elPrep 4 implements the semantics of GATK 4, which produces slightly different results from GATK 3.8.

We also set up a scaling experiment on Amazon Web Services, running the benchmark on a wide range of cloud servers that differ in terms of available RAM and CPUs (not shown here). These results indicate that elPrep 4 scales better than GATK 4. Because of this, the dollar cost to run elPrep remains stable when increasing the hardware resources to reduce the runtime.

References

[1] C Herzeel, P Costanza, D Decap, J Fostier, and W Verachtert. elPrep 4: A multithreaded framework for sequence analysis. *PLoS ONE*, 14(2), 2019.

[2] C Herzeel, P Costanza, D Decap, J Fostier, and J Reumers. elPrep: High-performance preparation of sequence alignment/map files for variant calling. *PLoS ONE*, 10(7), 2015.

[3] P Costanza, C Herzeel, and W Verachtert. A comparison of three programming languages for a full-fledged next-generation sequencing tool. *bioRxiv*, 2019.

Exploring the uncharacterized human proteome using neXtProt

Paula DUEK[1], Alain GATEAU[1], Amos BAIROCH[1,2] and Lydie LANE[1,2]

[1] CALIPHO Group, SIB-Swiss Institute of Bioinformatics, CMU, Michel-Servet 1, 1211 Geneva 4, Switzerland

[2] Department of Microbiology and Molecular Medicine, Faculty of Medicine, University of Geneva, CMU, Michel-Servet 1, 1211 Geneva 4, Switzerland

Corresponding Author: Lydie.lane@sib.swiss

Paper Reference: **Duek *et al.* (2018) Exploring the uncharacterized human proteome using neXtProt, J Proteome Res., 2018, https://doi: 10.1021/acs.jproteome.8b00537**

Keywords Functional annotation, human proteins, data mining, knowledge base, SPARQL

1. Introduction

Around 20,000 protein-coding genes have been predicted from the analysis of the human genome. In recent years, omics technologies allowed to collect a huge amount of data in terms of protein validation, protein–protein interactions, genetic variants, gene/protein expression and 3D structure, contributing to develop a more precise picture of the human proteome. neXtProt (www.nextprot.org) is a knowledge platform developed at the SIB Swiss Institute of Bioinformatics that aims to provide a state of the art representation of the knowledge about the human proteome by converting high quality data from a variety of heterogeneous sources into annotations with fully traceable evidence [1]. neXtProt's data model is based on RDF (Resource Description Framework), a core semantic web technology allowing to share and link data worldwide using common identifiers. Using the RDF query language (SPARQL), data can be retrieved not only from neXtProt but also from other semantically compatible databases.

neXtProt is used as reference knowledge base for the Human Proteome Project (HPP) from the Human Proteome Organization (HUPO), which aims to fill gaps in the knowledge of all human protein-coding genes [2]. Despite the accumulation of omics data, many human proteins are still partially functionally characterized and about 10% of them are devoid of any functional information. Although new technologies of targeted genome editing accelerate initial steps in protein characterization, understanding the function(s) of each protein in its biological context requires individual time-consuming studies, making the functional characterization of all human proteins a huge challenge. In the last five years, only 8-10 papers describing newly characterized human proteins were published each month. At this pace, the number of uncharacterized proteins will only decrease by 25% in the next five years. In order to speed up, 14 teams of the HPP consortium committed to initiate functional studies on such proteins using a variety of approaches and workflows [3].

The aim of our study was to explore the human "functionally dark proteome" using neXtProt and a combination of other resources in order to support such experimental characterization projects.

2. Results

The advanced, SPARQL-based search functionality of neXtProt was used to retrieve the human proteins that lack functional annotation. Despite the constant effort of curators to keep up to date with the literature, annotation gaps and delays are unavoidable. A systematic exploration of the available literature led to propose functional annotation updates for 113 proteins based on experimental reports.

This curation step led to the establishment of a consolidated list of 1,862 uncharacterized human proteins, among which 1,187 have been experimentally shown to exist in human biological samples.

The SPARQL-based search was also extensively used to explore the landscape of the uncharacterized human proteome in terms of subcellular locations, protein–protein interactions, tissue expression, association with diseases, and 3D structure. This information was complemented with tissue expression data from the Human

Protein Atlas [4], and phylogeny and phenotype data in model organisms from other resources cross-referenced by neXtProt.

Examining the collected information allowed us to propose functional hypotheses for 26 uncharacterized human proteins covering the fields of cilia biology, male reproduction, metabolism, nervous system, immunity, inflammation, RNA metabolism and chromatin biology. These hypotheses will require experimental validation in human and/or model organisms before they can be considered for annotation.

3. Conclusions and perspectives

This highlighted paper is an important contribution to the global community effort to fill the gaps in the functional annotation of the human proteome. The consolidated list of 1,862 uncharacterized human proteins is now used as a reference for the HPP functional characterization project. For all the proteins for which the function is unknown, generation of high quality experimental data and integration of this information in databases will be a key step toward an understanding of their role in the human body.

neXtProt will continue to collaborate with data providers and other bioinformatics resources to transform data into knowledge and provide tools to explore it. neXtProt data and query tool are open source and we will be glad to collaborate with other resources to enhance interoperability, improve the quality of functional predictions, and speed up the functional characterization of the human proteome.

Acknowledgements

This work was supported by SIB Swiss Institute of Bioinformatics and University of Geneva. The authors thank the neXtProt and HPP teams for their commitment in their projects, Lionel Breuza and Sylvain Poux from the UniProtKB/Swiss-Prot team for their valuable feedback, Nicolas Roggli for providing various scripts and help with parsing data.

References

1. Pascale Gaudet, Pierre-André Michel, Monique Zahn-Zabal, Aurore Britan, Isabelle Cusin, Marcin Domagalski, Paula Duek, Alain Gateau, Anne Gleizes, Valérie Hinard, Valentine Rech de Laval, Jin Lin, Frédéric Nikitin, Mathieu Schaeffer, Daniel Teixeira, Lydie Lane, Amos Bairoch. The neXtProt knowledgebase on human proteins: 2017 update. *Nucleic Acids Res*., 45(D1):D177-D182, 2017.
2. Gilbert S. Omenn, Lydie Lane, Christopher M. Overall, Fernando Corrales, Jochen M. Schwenk, Young-Ki Paik, Jennifer E. Van Eyk, Siqi Liu, Michael Snyder, Mark S. Baker, Eric W. Deutsch. Progress on identifying and characterizing the human proteome: 2018 metrics from the HUPO Human Proteome Project. *J Proteome Res*., in press, 2018.
3. Young-Ki Paik, Lydie Lane, Takeshi Kawamura, Yu-Ju Chen, Je-Yoel Cho, Joshua LaBaer, Jong Shin Yoo, Gilberto Domont, Fernando Corrales, Gilbert S. Omenn, Alexander Archakov, Sergio Encarnación-Guevara, Siqi Liu, Gashem Hosseini Salekdeh, Jin-Young Cho, Chae-Yeon Kim, Christopher M. Overall. Launching the C-HPP neXt-CP50 pilot project for functional characterization of identified proteins with no known function. *J Proteome Res.,* in press, 2018.
4. Mathias Uhlén, Linn Fagerberg, Björn M. Hallström, Cecilia Lindskog, Per Oksvold, Adil Mardinoglu, Åsa Sivertsson, Caroline Kampf, Evelina Sjöstedt, Anna Asplund, IngMarie Olsson, Karolina Edlund, Emma Lundberg, Sanjay Navani, Cristina Al-Khalili Szigyarto, Jacob Odeberg, Dijana Djureinovic, Jenny Ottosson Takanen, Sophia Hober, Tove Alm, Per-Henrik Edqvist, Holger Berling, Hanna Tegel, Jan Mulder, Johan Rockberg, Peter Nilsson, Jochen M. Schwenk, Marica Hamsten, Kalle von Feilitzen, Mattias Forsberg, Lukas Persson, Fredric Johansson, Martin Zwahlen, Gunnar von Heijne, Jens Nielsen, Fredrik Pontén, Tissue-based map of the human proteome. *Science* 347 (6220), 1260419, 2015.

How build up soil bacterial co-occurrence networks from wide spatial scale sampling?

Battle KARIMI[1], Samuel DEQUIEDT[1], Sébastien TERRAT[1], Claudy JOLIVET[2], Dominique ARROUAYS[2], Patrick WINCKER[3], Corinne CRUAUD[3], Antonio BISPO[2,4], Nicolas CHEMIDLIN-PREVOST BOURE[1], Lionel RANJARD[1]*.

[1] Agroécologie, AgroSup Dijon, INRA, Université Bourgogne Franche-Comté, F-21000 Dijon, France

[2] INRA Orléans - US 1106, Unité INFOSOL, Orléans - France

[3] CEA / Institut de Génomique / Genoscope, Evry cedex - France

Corresponding Author: battle.karimi@inra.fr

Paper Reference: **Karimi *et al.* (2019) Biogeography of Sol Bacterial Networks along a Gradient of Cropping Intensity, Genome Biology, 2019, 9:3812.** https://doi.org/10.1038/s41598-019-40422-y

Abstract *Although land use drives soil bacterial diversity and community structure, little information about the bacterial interaction networks is available. Here, we investigated bacterial co-occurrence networks in soils under different types of land use (forests, grasslands, crops and vineyards) by sampling 1798 sites in the French Soil Quality Monitoring Network covering all of France. An increase in bacterial richness was observed from forests to vineyards, whereas network complexity respectively decreased from 16,430 links to 2,046. However, the ratio of positive to negative links within the bacterial networks ranged from 2.9 in forests to 5.5 in vineyards. Networks structure was centered on the most connected genera (called hub), which belonged to Bacteroidetes in forest and grassland soils, but to Actinobacteria in vineyard soils. Overall, our study revealed that soil perturbation due to intensive cropping reduces strongly the complexity of bacterial network although the richness is increased. Moreover, the hub genera within the bacterial community shifted from copiotrophic taxa in forest soils to more oligotrophic taxa in agricultural soils.*

Keywords Soil – Nation wide scale - Bacterial Network – Hub Genera – Land use

Contrary to the macrobial communities, the ecological interaction network of microbial communities was viewed as impossible to investigate for long time. This was probably due to the lack of observability and measurement of the biotic interactions at the micro-scale but also to the little knowledge on the ways the microbes interact. Nevertheless, from the last 10 years, microbial ecology is flooded by network analysis. This phenomenon is related to both the advent of high throughput sequencing to study the microbial communities and the reborn of a statistical forgotten tool, the co-occurrence analysis. Although the co-occurrence of taxa was not strictly a biological interaction, the co-occurrence networks are considered as a satisfying approach to identify the potential microbio-sociological networks.

Most of studies in microbial ecology are based on oriented sampling, on a local spatial scale and with 3 to 10 field replicates, to reconstruct the co-occurrence network. The results from these studies are powerful to compare different treatments in specific environmental conditions, regarding the location and the date of sampling. Consequently, they are few generalizable to other environmental conditions, minimizing the scope of the conclusions. One way to provide strong generalized results is to use large scale, intensive and systematic sampling. Although this kind of sampling is efficient to provide several repetitions of networks and so the opportunity to test statistically the comparisons between modalities, it needs to develop one adapted methodology for network analysis.

In this paper, we investigated the soil bacterial co-occurrence network along a gradient of cropping intensity. We used the data from the French Network of Soil Quality Measurement (RMQS) which represents the most intensive soil sampling system on a wide spatial scale, due to its extensive area covered (5.5 10^5 km^2) and the high sampling resolution (about 2200 sites distributed along a systematic grid). From this sampling, the soil bacterial communities' of 1798 sites have been successfully characterized using 454-

pyrosequencing of 16S rRNA gene. Four land uses have been mainly found across the sampling grid: forest, grassland, crop system, vineyards/orchards with respectively 492, 464, 740 and 36 soils. Thus, the structure of bacterial co-occurrence networks has been compared between these four land uses. After multiple tests to measure the heterogeneity across samples and to set the different methodological parameters, we computed 100 repetitions of network for each land use, each based on a subset of 25 sites randomly selected. The co-occurrences and their significance were established using a spearman correlation coefficient, validated by a False-discovery Rate (also named Benjamin-Hochberg) correction. To compare the structure of networks between land uses, 6 metrics were computed: the number of links, the connectance, the average path length, the average degree, and the ratio between positive and negative links. Moreover, the hub taxa were identified and statistically compared between land uses.

Visual comparison of the networks for each land use revealed a significant shift in structure ranging from a highly connected, tightly closed structure for forests to a sparse, open structure for vineyards (Fig 1). The bacterial networks in forest soils formed a condensed cluster. In grassland soils, the cluster seemed to split into two parts with several long chains extending from the clusters. In soils under crop systems, one of clusters split into several large satellites which remained connected. In vineyards soils, many of the links seemed to be lost and the satellites were smaller and less inter-connected. Numerically, the network complexity respectively decreased from 16,430 links in forests to 2,046 in vineyards/orchards (decreasing of 87%) whereas the bacterial richness was increasing from forests to vineyards. However, the ratio of positive to negative links within the bacterial networks ranged from 2.9 in forests to 5.5 in vineyards. Networks structure was centered on the most connected genera (called hub), which belonged to Bacteroidetes in forest and grassland soils, but to Actinobacteria in vineyard soils.

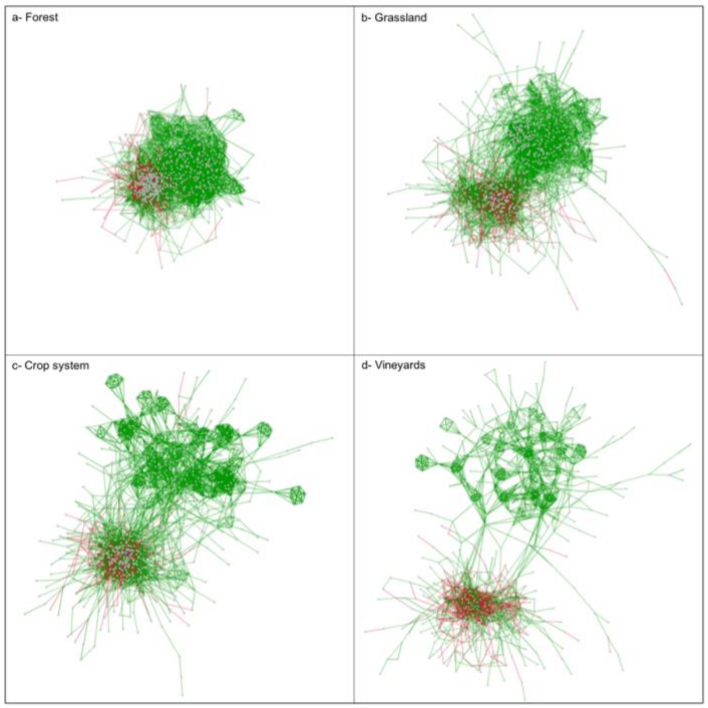


Figure 1 Visualization of the most complex network among the 100 replicates for the 4 land uses. The red edges represent the negative links and the green edges represent the positive links. The most complex network was the one with the most links, the highest connectance and the highest average degree.

Altogether our study demonstrated that soil bacterial co-occurrence networks are different between land use types and are strongly shaped by the cropping intensity. We hypothesized that changes in the bacterial network would occur mainly in response to shifts in the heterogeneity and connectivity of the mosaic of microbial habitats as well as to the availability of C-substrates. Beyond the classical information obtained from bacterial richness or whole taxonomic composition, co-occurrence network analysis provides complementary insights into biotic interactions and niche connectivity, which could have repercussions on community stability and soil functioning. Finally, the use a wide scale sampling has allowed robust analysis to provide a generalized conclusion about the effect of cropping intensity on the soil microbial communities.

References

All references are detailed in the corresponding paper.

Merging of phenotypic information from cytometric profiles at the single-cell resolution.

Adrien Leite Pereira[1], Olivier Lambotte[1], Roger Le Grand[1], Antonio Cosma[1], and Nicolas Tchitchek*[1]

[1]Research Center for Immunology of Viral Infections and Autoimmune Diseases – Commissariat à l'Énergie Atomique et aux Énergies Alternatives - CEA, Institut National de la Santé et de la Recherche Médicale - INSERM : U1184, Université Paris Sud - Paris XI – France

Résumé

Background: Flow and mass cytometry are experimental techniques used to measure the level of proteins expressed by cells at the single-cell resolution. Several algorithms were developed in flow cytometry to increase the number of simultaneously measurable markers. These approaches aim to combine phenotypic information of different cytometric profiles obtained from different cytometry panels.
Results: We present here a new algorithm, called CytoBackBone and recently published in Bioinformatics, which can merge phenotypic information from different cytometric profiles. This algorithm is based on nearest-neighbor imputation, but introduces the notion of acceptable and non-ambiguous nearest neighbors. We demonstrated that the merging results produced by CytoBackBone are symmetrical and more-stringent compared to other approaches. Mass cytometry data were used to illustrate the merging of cytometric profiles obtained by the CytoBackBone algorithm.
Impact: In principle, there is no limit to the number of cytometric profiles that can be merged by the CytoBackBone algorithm. This new algorithm will be key to improve the depth of the cell phenotyping in immunological studies. We are currently using this algorithm to characterize inflammation in HIV-ART patients based on merged cytometric profiles of 72 cell markers.

*Intervenant

Sequana coverage: detection and characterization of genomic variations using running median and mixture models

Thomas Cokelaer[1,2]

[1] Institut Pasteur – Platform Biomics – 25-28 Rue du docteur Roux, Paris, France.

[2] Institut Pasteur – Bioinformatics and Biostatistics Hub – C3BI, USR 3756 IP CNRS – Paris, France

Corresponding Author: thomas.cokelaer@pasteur.fr

Paper Reference: **Desvillechabrol et al. (2018) Sequana coverage: detection and characterization of genomic variations using running median and mixture models, Giga Science, Volume 7, Issue 12, Decembre 2018, giy110, https://doi.org/10.1093/gigascience/giy110**

Abstract *We provide a stand-alone application, sequana_coverage, that reports genomic regions of interest (ROIs) that are significantly over- or underrepresented in high-throughput sequencing data. Significance is associated with the events as well as characteristics such as length of the regions. The algorithm first detrends the data using an efficient running median algorithm. It then estimates the distribution of the normalized genome coverage with a Gaussian mixture model. Finally, a z-score statistic is assigned to each base position and used to separate the central distribution from the ROIs (i.e., under- and overcovered regions). HTML reports provide a summary with interactive visual representations of the genomic ROIs with standard plots and metrics. Genomic variations such as single-nucleotide variants or CNVs can be effectively identified at the same time.*

Keywords genome coverage; sequencing depth; running median; Sequana; NGS; Python; Snakemake; CNV

1. Introduction

In addition to mapping quality information, the genome coverage contains valuable biological information such as the presence of repetitive regions, deleted genes, or copy number variations (CNVs). It is essential to take into consideration atypical regions, trends (e.g., origin of replication), or known and unknown biases that influence coverage. It is also important that reported events have robust statistics (e.g. z-score) associated with their detections as well as precise location.

Figure 1: Genome coverage (bacterial genome) with under of over covered regions.

In Figure 1 we show a typical example of genome coverage with an origin of replication effect and a gene deletion that affects the estimation of the mean coverage along the genome. In order to account for those effects and automatically detect all over and under covered regions, we design a tool integrated into the sequana library (https://sequana.readthedocs.io). The main difference with existing tools is the ability to detect narrow events (a few bases) as well as large events (tens of kb).

2. Results

In this work we describe a novel approach that can efficiently detect various types of genomic regions. The algorithm does not target any specific type of genomic variations but instead systematically reports all positions (with a z-score) that have depth departing from the overall distribution. The algorithm normalizes the genome coverage using a running median and then calculate a robust statistic (z-score) for each base position based on the parameter estimation of the underlying distribution. This allows us to obtain robust and non-constant thresholds at each genome position.

Figure 2: Normalised genome coverage using a running median estimator

In the first part of the talk, we describe the proposed novel method of detecting under or over represented regions in the genome coverage data. In particular, we describe (i) the running median used to detrend the genome coverage, (ii) the statistical methods used to characterize the central distribution from which outliers can be identified and (iii) a double thresholds method proposed to cluster the ROIs.

Figure 3: Detection of a depleted region. CNVnator (thick yellow seg-ments) and sequana coverage (thin green segments and dots) identifies the 6,300 long event with the correct location as well as shorter events (500bp) but sequana_coverage also identifies the other short events (few bases long).

In the second part, we present an application for CNV detections. In particular, in the context of bacterial genomes, we show how this implementation out-performs some established tools in not only detecting CNVs but also precisely identifies their location and number. As a test example, we use 6 isolates of Staphylococcus aureus. We describe the different between our implementation and two established tools namely CNVnator and CNOGpro, the latter being dedicated to the detection of CNV in bacterial genomes.

References

1. Goodwin S, McPherson JD, McCombie WR. Coming of age: ten years of next-generation sequencing technologies. Nat Rev Genet 2016; 17 (6):333–351
2. Lander ES, Waterman MS. Genomic mapping by finger-printing random clones: a mathematical analysis. Genomics 1988; 2 (3):231–239.
3. The Sequana resources hithub repository https://github.com/sequana/resources/coverage
4. Abyzov A, Urban AE, Snyder M, et al. CNVnator: an approach to discover, genotype, and characterize typical and atypical CNVs from family and population genome sequencing. Genome Res 2011; 21, 974–984
5. Cokelaer T, Desvillechabrol D, Legendre R, et al. Sequana: a set of Snakemake NGS pipelines. Journal of Open Source Software 2017; 2 ,16 https://doi.org/10.21105/joss.00352.
6. Brynildsrud O, Snipen LG, Bohlin J. CNOGpro: detection and quantification of CNVs in prokaryotic whole-genome sequencing data. Bioinformatics 2015; 31 (11):1708–1715.

Cyclin A2 and E1 genomic alterations define a specific subclass of hepatocellular carcinoma subclass with a rearrangement signature of replication stress

Quentin BAYARD[1], Léa MEUNIER[1]*, Camille PENEAU[1]*, Victor RENAULT[2], Jayendra SHINDE[1], Jean-Charles NAULT[1,3,4], Iadh MAMI[1], Gabrielle COUCHY[1], Giuliana AMADDEO[5,6], Emmanuel TUBACHER[2], Delphine BACQ[7], Vincent MEYER[7], Tiziana La BELLA[1], Audrey DEBAILLON-VESQUE[8], Paulette BIOULAC-SAGE[9,10], Olivier SEROR[1,11], Jean-Frédéric BLANC[8,9], Julien CALDERARO[5,12], Jean-François DELEUZE[2,7], Sandrine IMBEAUD[1], Jessica ZUCMAN-ROSSI[1,13]** & Eric LETOUZÉ[1]**

1 Centre de Recherche des Cordeliers, INSERM, Sorbonne Université, USPC, Université Paris Descartes, Université Paris Diderot, Functional Genomics of Solid Tumors (FunGeST), F-75006 Paris, France
2 Laboratory for Bioinformatics, Fondation Jean Dausset – CEPH, Paris, France.
3 Liver unit, Hôpital Jean Verdier, Hôpitaux Universitaires Paris-Seine-Saint-Denis, Assistance-Publique Hôpitaux de Paris, APHP, Bondy, France.
4 Unité de Formation et de Recherche Santé Médecine et Biologie Humaine, Université Paris13, Communauté d'Universités et Etablissements Sorbonne Paris Cité, Paris, France.
5 Inserm, U955, Team 18, Université Paris-Est Créteil, Faculté de Médecine.
6 Assistance Publique-Hôpitaux de Paris, Service d'Hépatologie, CHU Henri Mondor.
7 Centre National de Recherche en Génomique Humaine, CEA, Evry, France.
8 Service Hépato-Gastroentérologie et Oncologie Digestive, Hôpital Haut-Lévêque, Centre Hospitalier Universitaire de Bordeaux, Bordeaux, France.
9 Université Bordeaux, Bordeaux Research in Translational Oncology, Bordeaux, France
10 Service de Pathologie, Hôpital Pellegrin, Centre Hospitalier Universitaire de Bordeaux, Bordeaux, France.
11 Radiology Department, Jean Verdier Hospital, Bondy, Hôpitaux Universitaires Paris-Seine-Saint-Denis, APHP, Bondy, France.
12 Assistance Publique-Hôpitaux de Paris, Département de Pathologie, Hôpital Henri Mondor, Créteil, France.
13 Assistance Publique-Hôpitaux de Paris, Hopital Européen Georges Pompidou, F-75015 Paris, France
Corresponding Author: J. Zucman-Rossi (jessica.zucman-rossi@inserm.fr) or E. Letouzé (eric.letouze@inserm.fr)

Paper Reference: **Bayard *et* al. (2018). Cyclin A2 and E1 genomic alterations define a specific subclass of hepatocellular carcinomas subclass with rearrangement signature of replication stress. Nature communication, 2018 Dec 7;9(1):5235. doi: 10.1038/s41467-018-07552-9.**

Abstract

Cyclins A2 and E1 regulate the cell cycle by promoting S phase entry and progression. Here, we identify a hepatocellular carcinoma (HCC) subgroup exhibiting cyclin activation through various mechanisms including hepatitis B virus (HBV) and adeno-associated virus type 2 (AAV2) insertions, enhancer hijacking and recurrent CCNA2 fusions. Cyclin A2 or E1 alterations define a homogenous entity of aggressive HCC, mostly developed in non-cirrhotic patients, characterized by a transcriptional activation of E2F and ATR pathways and a high frequency of RB1 and PTEN inactivation. Cyclin-driven HCC display a unique signature of structural rearrangements with hundreds of tandem duplications and templated insertions frequently activating TERT promoter. These rearrangements, strongly enriched in early-replicated active chromatin regions, are consistent with a break-induced replication mechanism. Pan-cancer analysis reveals a similar signature in BRCA1-mutated breast and ovarian cancers. Together, this analysis reveals a new poor prognosis HCC entity and a rearrangement signature related to replication stress.

Keywords

Cancer genomics

Whole Genome sequencing

Structural rearrangement Signature

Signature analysis of Structural Variants reveals a new subclass of hepatocellular carcinoma characterized by Cyclin A2/E1 alterations

Quentin BAYARD[1], Léa MEUNIER[1]*, Camille PENEAU[1]*, Benedict Monteiro[1], Victor RENAULT[2], Jayendra SHINDE[1], Jean-François DELEUZE[2], Sandrine IMBEAUD[1], Jessica ZUCMAN-ROSSI[1,3]** & Eric LETOUZE[1-4]**

1 Centre de Recherche des Cordeliers, INSERM, Sorbonne Université, USPC, Université Paris Descartes, Université Paris Diderot, Functional Genomics of Solid Tumors (FunGeST), F-75006 Paris, France
2 Laboratory for Bioinformatics, Fondation Jean Dausset – CEPH, Paris, France.
3 Assistance Publique-Hôpitaux de Paris, Hopital Européen Georges Pompidou, F-75015 Paris, France

Corresponding Author: E. Letouzé (eric.letouze@inserm.fr)

Bayard *et* al. (2018). Cyclin A2 and E1 genomic alterations define a specific subclass of hepatocellular carcinomas subclass with rearrangement signature of replication stress. Nature communication, 2018 Dec 7;9(1):5235. doi: 10.1038/s41467-018-07552-9.

Abstract

Hepatocellular-Carcinoma (HCC) is the 3rd cause of cancer death worldwide. Liver carcinogenesis is the result of a complex multistep process generated by pro-oncogenic genetic alterations. Structural variants (deletions, duplications, inversions, translocations of chromosomic regions, or more complex events) occur in tumor cells due to various biological processes, related to DNA replication, and fuel tumorigenesis by activating oncogenes or disrupting tumor suppressors.

Whole-Genome Sequencing (WGS) followed by Structural Variant (SV) calling highlighted that some tumors display high numbers of rearrangements with a very specific pattern. This can reveal a common underlying biological defect, illustrated by the duplicator phenotypes in *BRCA1* and in *CDK12*-altered breast [1], ovarian [2] and prostate cancers [3]. However, such phenotypes have not been explored in liver cancer so far. We analyzed SVs across 354 liver cancer genomes (LiC1138, n=49; TCGA, n=257; ICGC, n=48) and used non-Negative Matrix Factorization (NMF) [4] to deconvolute Rearrangement Signatures of biological processes from catalogs of SVs among each samples in the cohort. This innovative approach allowed us to identify 6 signatures, operative at low levels in most tumors but highly active in small tumor subgroups showing extreme structural rearrangement phenotypes. More specifically, we describe a new signature (RS1) of small tandem duplications and Templated Insertion Cycles (T.I.C), a complex mechanism involving inter-connected chains of chromosome translocations, associated with a specific subgroup of aggressive HCC exhibiting cyclin A2 or E1 activation through various mechanisms including HBV and AAV2 viral insertions, enhancer hijacking and recurrent CCNA2 fusions (CCN-HCC). In these tumors, premature S phase entry leads to intense replication stress and generates hundreds of focal structural rearrangements. These rearrangements, strongly enriched in early-replicated active chromatin regions, are consistent with a break-induced replication mechanism. We finally used a negative binomial regression to model the uneven breakpoint distribution of RS1 events in order to highlight hotspot of "second hits" driver alterations. Of note, TERT activating promoter rearrangements were significantly enriched in CCN-HCC.

In conclusion, signature analysis of SVs allowed us to define a CCN-HCC phenotype, which corresponds to 7% of our HCC series, and defines a homogenous entity of aggressive tumors, mostly developed in non-cirrhotic patients without classical risk factors that may benefit from therapies targeting replication stress.

References

[1] S. Nik-Zainal *et al.*, "Landscape of somatic mutations in 560 breast cancer whole-genome sequences," *Nature*, vol. 534, no. 7605, pp. 47–54, 2016.
[2] T. Popova *et al.*, "Ovarian cancers harboring inactivating mutations in CDK12 display a distinct genomic instability pattern characterized by large tandem duplications," *Cancer Res.*, 2016.
[3] D. A. Quigley *et al.*, "Genomic Hallmarks and Structural Variation in Metastatic Prostate Cancer," *Cell*, 2018.
[4] L. B. Alexandrov, S. Nik-Zainal, D. C. Wedge, P. J. Campbell, and M. R. Stratton, "Deciphering Signatures of Mutational Processes Operative in Human Cancer," *Cell Rep.*, 2013.

Une nouvelle méthode pour évaluer l'impact des mesures de similarité sémantique sur l'annotation d'un groupe de gènes

Aarón Ayllón-Benítez[1,2], Fleur Mougin[1,2] et Patricia Thébault[2]
[1] Univ. Bordeaux, Inserm UMR 1219, Bordeaux Population Health Research Center, ERIAS team, France
[2] Univ. Bordeaux, CNRS UMR 5800, LaBRI, France

Auteur référent: aaron.ayllon-benitez@u-bordeaux.fr

Article présenté: **Ayllón-Benítez *et al.* (2018) A new method for evaluating the impacts of semantic similarity measures on the annotation of gene sets. doi:10.1371/journal.pone.0208037**

Résumé *Les méthodes statistiques basées sur l'enrichissement permettent de comprendre le(s) processus biologique(s) dans le(s) quel(s) un groupe de gènes différentiellement exprimés est impliqué. Cependant, ces méthodes tendent à mettre l'accent sur les gènes les plus étudiés et affecter l'interprétation des données biologiques, en négligeant les gènes pour lesquels les connaissances (ou annotations) sont encore en évolution. Des approches alternatives existent pour résoudre ces problèmes, notamment celles exploitant des mesures de similarité sémantique. Cet article présente ainsi une nouvelle méthode qui analyse l'impact de différentes mesures de similarité sémantique pour annoter un groupe de gènes. Plusieurs mesures de similarité exploitant différents types de connaissances de la* `Gene Ontology` *ont été considérées pour annoter un groupe de gènes de manière synthétique. Les mesures de similarité basées sur les connaissances liées aux nœuds ont généré de meilleurs résultats comparativement à celles basées sur les arêtes.*

Mots-clés Similarité sémantique, Gene Ontology, Annotation d'un groupe de gènes

1 Introduction

Les avancées dans l'analyse de l'expression différentielle de gènes a suscité un vif intérêt pour l'étude des groupes de gènes qui partagent une similarité d'expression dans un même processus biologique. Les approches classiques pour interpréter l'information biologique reposent sur l'utilisation de méthodes statistiques d'enrichissement. Cependant, ces méthodes tendent à mettre l'accent sur les gènes les plus connus tout en générant de la redondance d'information par la non prise en compte des possibles relations entre les termes d'annotation. Une alternative consiste à utiliser les mesures de similarité sémantique pour regrouper les termes d'annotation selon leur similarité et faciliter ainsi l'interprétation du groupe de gènes étudié. L'article [1] analyse l'impact de différentes sortes de mesures de similarité sémantique avec l'objectif de proposer une annotation synthétique d'un groupe de gènes donné.

2 Matériels et Méthodes

La `Gene Ontology` (GO) est une ressource décrivant les processus et fonctions des produits de gènes dans le monde du vivant. La structure de GO est constituée d'un graphe orienté acyclique qui compte plus de 44 000 termes connectés par différents types de relations (*e.g.*, *is_a*, *part_of*, *regulates*). Les annotations gène - terme GO sont obtenues à partir de la base de données GO Annotation (GOA).

A partir de l'ensemble des termes GO extraits pour chaque gène du groupe d'intérêt, un premier filtre a été appliqué pour supprimer les annotations n'apportant pas d'information pertinente pour le gène (*i.e.*, annotations redondantes ou incomplètes).

Afin d'étudier l'impact des différentes mesures de similarité sémantique, nous en avons sélectionné parmi les trois classes suivantes, définies par Pesquita *et al.* [2] : mesures basées sur les nœuds, mesures basées sur les arêtes et mesures hybrides. Les mesures basées sur les nœuds calculent la similarité entre deux termes à partir des propriétés spécifiques aux termes, comme leur profondeur ou leur contenu d'information (CI). Les mesures de type basées sur les arêtes exploitent la distance qui sépare deux termes GO au sein du graphe GO. Les mesures hybrides utilisent, quant à elles, une combinaison des deux types précédents.

Ensuite, nous avons examiné la capacité de chaque mesure de similarité sémantique à obtenir les meilleures partitions des termes d'annotation en évaluant la pertinence des partitions du clustering et l'impact des différentes méthodes hiérarchiques de clustering.

Pour identifier les termes les plus synthétiques du groupe de gènes tout en évaluant la combinatoire des solutions, nous avons développé un algorithme de parcours de graphe afin de récupérer tout d'abord un ou plusieurs termes (nombre dépendant de la taille du cluster) représentatifs pour chaque cluster de termes obtenu. À partir des solutions résultantes, nous avons examiné l'efficacité de chaque mesure de similarité sémantique pour (i) réduire le nombre de termes d'annotation tout en sélectionnant les termes les plus représentatifs du groupe de gènes et (ii) fournir une annotation synthétique incluant le plus de gènes possibles. La combinaison de ces deux critères, essentiellement quantitatifs, a permis d'estimer la capacité de chaque mesure de similarité sémantique à produire une annotation plus pertinente et synthétique pour un groupe de gènes donné.

3 Résultats et Discussion

Nous avons étudié l'impact des mesures de similarité sémantique en utilisant deux jeux de données de l'organisme "homo sapiens" qui contiennent respectivement 260 et 360 groupes de gènes liés à la réponse immunitaire dans le cadre de diverses maladies. Les différentes évaluations sur les partitions de clustering ont montré de meilleurs résultats avec les mesures basées sur les nœuds par rapport à celles basées sur les arêtes et de mauvais résultats pour les mesures hybrides.

Pour étudier l'impact de chaque mesure de similarité sémantique sur l'obtention d'une annotation synthétique, nous avons analysé la quantité de termes retenus et le nombre de gènes couverts en observant la pertinence en terme d'information biologique (mesure basée sur le CI). En comparant avec la méthode d'enrichissement DAVID [3], nous avons montré que notre approche donne de meilleurs résultats avec la majorité des mesures de similarité sémantique, les mesures basées sur les nœuds étant les meilleures. Les mesures de similarité sémantique permettent ainsi de trouver un bon équilibre pour garantir la meilleure couverture du nombre de gènes avec un nombre minimum de termes (tout en gardant une information pertinente et synthétique).

4 Conclusion

Les principaux problèmes qui se posent dans la recherche de signatures génétiques sont liés à l'étude de la fonction biologique des groupes de gènes. Ces objectifs peuvent être facilement atteints en utilisant des méthodes d'enrichissement telles que DAVID. Cependant, ces méthodes présentent des limitations impliquant une perte d'information et une redondance d'information inutile. Pour répondre à ces limitations, la bioinformatique propose diverses stratégies allant de l'analyse de l'enrichissement aux mesures de similarité sémantique. Ces dernières approches ont fait l'objet de nombreuses études de la part de la communauté scientifique afin de fournir un large éventail de mesures. Bien que ces mesures soient souvent combinées à des méthodes d'enrichissement, leur utilisation *a priori* peut avoir une grande incidence sur l'interprétation des ensembles de données biologiques. Dans ce cadre, nous avons élaboré une approche qui utilise des mesures de similarité sémantique afin de réaliser une interprétation robuste. Nous avons choisi de centrer notre analyse sur neuf mesures couvrant diverses caractéristiques des termes GO et exploré leurs avantages et leurs inconvénients pour fournir des informations pertinentes aux experts du domaine. Ainsi, d'un point de vue biologique, notre cadre analytique a permis d'analyser leur capacité à synthétiser l'information et à fournir le meilleur compromis possible pour conserver les informations pertinentes.

Références

[1] A. Ayllón-Benítez, F. Mougin, J. Allali, R. Thiébaut, and P. Thébault. A new method for evaluating the impacts of semantic similarity measures on the annotation of gene sets. *PLoS One*, 13(11), 2018.

[2] C. Pesquita, D. Faria, A.O. Falcão, P. Lord, and F.M. Couto. Semantic similarity in biomedical ontologies. *PLoS Comput. Biol*, 5(7), 2009.

[3] G. Dennis, B.T. Sherman, D.A. Hosack, J. Yang, W. Gao, H.C. Lane, and R.A. Lempicki. David : database for annotation, visualization, and integrated discovery. *Genome Biol*, 4(9), 2003.

A De Novo Robust Clustering Approach for Amplicon-Based Sequence Data

Alexandre Bazin*[1] Didier Debroas[2]

Engelbert Mephu Nguifo[3]

[1]contact@alexandrebazin.com

University Clermont Auvergne, CNRS, LIMOS, F-63000

CLERMONT-FERRAND, FRANCE

[2]didier.debroas@uca.fr

University Clermont Auvergne, CNRS, LMGE, F-63000

CLERMONT-FERRAND, FRANCE

[1]engelbert.mephu_nguifo@uca.fr

University Clermont Auvergne, CNRS, LIMOS, F-63000

CLERMONT-FERRAND, FRANCE

Abstract

When analyzing microbial communities, an active and computational challenge concerns the categorization of 16S rRNA gene sequences into operational taxonomic units (OTUs). Established clustering tools use a one pass algorithm in order to tackle high numbers of gene sequences and produce OTUs in reasonable time. However, all of the current tools are based on a crisp clustering approach, where a gene sequence is assigned to one cluster. The weak quality of the output compared to more complex clustering algorithms, forces the user to post-process the obtained OTUs. Providing a membership degree when assigning a gene sequence to an OTU, will help the user during the post-processing task. Moreover it is possible to use this membership degree to automatically evaluate the quality of the obtained OTUs. So the goal of this work is to propose a new clustering approach that takes into account uncertainty when producing OTUs, and improves both the quality and the presentation of the OTUs results.

1 Introduction

Studying the structure of the communities in an ecosystem is central in environmental microbiology Hugoni et al. (2013); Roux et al. (2011). The biosphere's diversity can be determined by amplifying and sequencing specific phylogenetic markers (e.g. 16S rRNA). From there, these amplicons need to be clusterized in "species" named Operational Taxonomic Units (OTUs) Chen et al. (2013); Li et al. (2012); Mahé et al. (2014); Westcott and Schloss (2015). As the volume of sequences has drastically increased in recent times, new clustering tools have emerged to treat the data in reasonable time. The currently used algorithms are, from the point of view of algorithmic complexity, the fastest available that do not produce random results. However, due to their simplicity, the reliability of the results are often discussed. These tools being essentially black boxes, their sensitivity to the sequence order, clustering threshold and structure of the data makes it that the users have no way of knowing whether better Operational Taxonomic Units (OTUs) could have been obtained with different parameters or even whether they correctly represent the data. In these circumstances, there is no choice but to blindly trust them.

Distance-based greedy clustering algorithm such as the ones implemented in OTUclust Albanese et al. (2015), VSEARCH Rognes et al. (2016), CD-HIT Li and Godzik (2006) or USEARCH Edgar (2010) all share the same base algorithm as shown in Algorithm 1.

While more sophisticated algorithms Antoine et al. (2014); Gath and Geva (1989); Pérez-Suárez et al. (2013); Hariz et al. (2006); Antoine et al. (2012) could produce better results quality-wise, their runtime would render them unusable on millions of sequences. As the quality of the OTUs is important, we have to find a way to improve it without increasing the runtime. The different available implementations use a variety of heuristics to counterbalance the simplicity of the algorithm but, to the best of our knowledge, no approach has tried to add a measure of uncertainty to the process. This is why, in order to help increase the quality and trustworthiness of the clustering, we propose to add uncertainty to this simple algorithm through the use of fuzzy clustering.

2 Method

2.1 Motivation

Distance-based greedy clustering algorithms, such as the one in VSEARCH, produce a number of OTUs and assign each sequence to one of them. The OTU to which a sequence is said to belong to is usually the first one to be encountered that is sufficiently close, i.e. within the specified threshold. This creates two problems :

- A sequence can only belong to a single OTU
- An OTU either includes or does not include a sequence

Having a sequence associated to a single OTU is expected as the ultimate output of the algorithm. For this reason, algorithms can stop after finding the first OTU that is close enough to a sequence, which speeds the computation up. However, not considering all the OTUs a sequence could be assigned to increases the sensitivity to the order - a weakness of these algorithms - and reduces the quality of the clustering. Indeed, what if two different OTUs are close enough ? Giving priority to the first generated OTU only creates a bias that no heuristic - such as sorting the sequences - could hope to overcome.

Moreover, by using strict thresholds, it is possible to have two nearly identical sequences such that one belongs to a particular OTU while the other does not. This strictness makes it so an OTU partitions the set of sequences into two sets inside of which sequences are

considered the same regardless of their distance to the center of the OTU. This lack of distinction between sequences that are isolated and sequences on the border of OTUs hides information that could help understand the data.

While these would not be problems were the clustering optimal, the need for fast algorithms gives rise to results that are not always trustworthy. The OTUs being presented as absolute, the end user has no choice, should consider them correct and cannot know whether the algorithm has encountered ambiguity. We believe that being less strict in the way the OTUs partition sequences would help produce better results from the end user's point of view.

2.2 Fuzzy Clustering

To help increase the quality of the clustering and maximize the information that can be gathered from the data, we propose to add uncertainty to the clustering by means of fuzzy sets.

We define a membership function $f_C(S)$ that, for an OTU C, associates a membership value to a sequence S. Usually, this value is either 0 or 1. Here, we propose to have $f_C(S)$ take its value in $\{\frac{n}{10} \mid n = 0 \dots 10\}$. This value represents the degree of membership and, as such, 1 means that the sequence **certainly** belongs to the OTU while 0 means that the sequence **certainly** does not belong to it. Other values represent uncertainty and are used to express that the sequence **nearly** belongs to the OTU. This membership value can

easily be computed from the distance between the sequence and the center of the OTU using two thresholds t_1 and t_2 such that $t_1 \geq t_2$. If the distance is less than the threshold t_1, the membership value is 1. If the distance is greater than t_2 the value is 0. If the distance is between t_1 and t_2, it increases gradually. Fuzzy OTUs are depicted in Fig. 1.

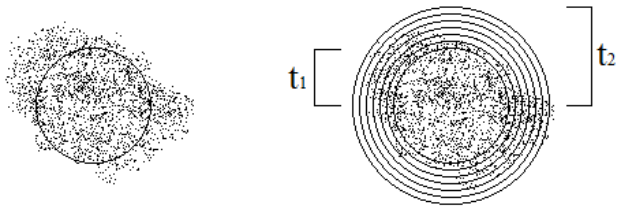


Figure 1: Representations of a Crisp (Left) and a Fuzzy (Right) OTUs.

Using fuzzy OTUs allows us to discern the difference between sequences close to the OTU and sequences extremely far. Using the parameters t_1 and t_2, we can tune the "detection radius" around OTUs to gather information that would normally be discarded by the clustering algorithm.

2.3 Evaluating fuzzy OTUs

Having a non-binary membership function produces OTUs that partition the sequences into multiple sets. If we consider only the sequences that belong (more or less) to an OTU, the repartition of their membership values provides information on the topology of the OTU. An

	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1
OTU1	6	4	1	1	0	3	8	13	29	88
OTU2	70	41	30	41	34	19	11	6	5	16

Table 1: Two example OTUs with the number of sequences that belong to them with each possible membership value.

ideal OTU would contain only sequences with a membership value of 1, meaning a group of sequences has been perfectly regrouped with a good threshold and no sequence lies ambiguously on the border. More realistically, a good OTU would contain many sequences with high membership values and little sequences with low values. A bad OTU with the majority of its sequences having low membership values could mean that the algorithm has chosen as a center a sequence on the border of a group or, even worse, between two distinct groups. Examples of such good and bad OTUs are given in Table 1.

We can quickly evaluate the quality of an OTU with this repartition. If we suppose that each sequence lowers the quality of the OTU depending on its membership value, we can use the following formula:

$$Quality(OTU) = 1 - \sum_{i=1}^{9} \omega_i \times \frac{\#\text{ sequences with membership value } i \times 0.1}{\#\text{ sequences in the OTU}}$$

with ω_i being the "cost" of having a sequence with membership value $i \times 0.1$. In our previous examples, and with the following values of ω_i

ω_1	ω_2	ω_3	ω_4	ω_5	ω_6	ω_7	ω_8	ω_9
1	0.9	0.8	0.7	0.6	0.5	0.4	0.3	0.2

Table 2: Example of weight values.

we obtain a quality of respectively 0.71 and 0.26 for OTU1 and OTU2, showing OTU1 is better.

A problem arises with singletons that always have perfect quality but these can safely be treated separately.

2.4 Choosing an OTU

A sequence can belong to multiple OTUs due to fuzzy membership. However, in the end, we want each sequence to be assigned to a single OTU. Hence, we have to choose one of the possible OTUs. We have two types of values left from the clustering process : membership and quality. The first one is based on the distance between the OTU and the sequence and the second one is used to recognize bad OTUs. Choosing the OTU with the best membership value is akin to running VSEARCH. Choosing the OTU with the best quality tends to create bigger OTUs that absorb distant sequences. To better compromise, we can use a linear combination of both values :

$$\alpha \times quality + \beta \times membership$$

Increasing the importance of the quality reduces the number of

OTUs containing sequences. When α is low, the "best" OTUs quality-wise absorb very close sequences that would have been attributed to other OTUs. When α gets too high, the best OTUs start absorbing all the sequences around them, effectively acting like an increase of the distance threshold.

2.5 Identifying ambiguous sequences

Distance-based greedy algorithms are good at clustering objects that are easy to cluster. Groups of very similar sequences that are different from the rest of the dataset are supposed to birth a new OTU while isolated singletons should be identified to be either removed or treated separately. A problem arises when groups of sequences are close to each other but not enough to be the same OTU. In this case and supposing the algorithm ideally chooses the centers of the OTUs, sequences can lie just between these OTUs. In the current implementations, these ambiguous sequences that must be assigned are usually put in OTUs of their own, increasing the number of OTUs and reducing the overall quality of the clustering.

Using fuzzy clustering allows us to identify these ambiguous sequences such as those depicted in Fig. 2. Using the previously mentioned choice strategy, they can be assigned to a good OTU even though they lie slightly outside of the distance threshold. However, their ambiguousness may be significant for the user. It is thus important to highlight their existence and the various fuzzy OTUs they

could have alternatively been assigned to.

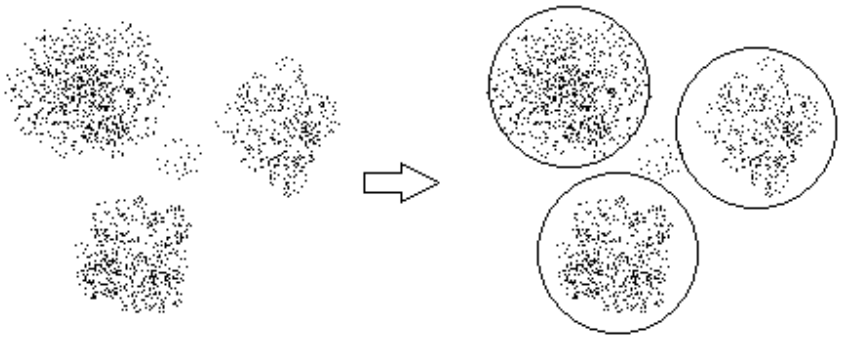


Figure 2: A Case of Ambiguous Sequences

3 Experimental Results

3.1 Relevant Metrics

Five criterion are important when evaluating the efficiency of a clustering method in the setting of environmental microbiology. First of all, the computation (**Time**) should be as fast as possible and the memory (**Memory**) usage should be low. Comparing runtimes with those of VSEARCH, that gives satisfactory results, is enough. The quality of the clustering itself is evaluated using the number of OTUs (**#OTUs**), singletons (**#singletons**) and pairs (**#pairs**). The way to use these three values to study the population of microorganisms is outside the scope of this work but it is important to note that they should be minimised and that the proportion of singletons should be as small as possible. Additionally, the average taxonomic distance (**Distance**) between sequences in a same OTU is used to represent the difference between the computed result and reality. The distance between two sequences in the taxonomy is defined as the sum of the lengths of the path from their nearest commonality. For example, if a sequence is classified as "bacteria;proteobacteria;betaproteobacteria" and the other is classified as "bacteria;proteobacteria;alphaproteobacteria ", their distance is 2 as each of them is at a distance 1 from their commonality ""bacteria;proteobacteria".

The comparison with VSEARCH is done using identical parameters when applicable.

3.2 Data and Results

We tested our algorithm on a dataset containing 100000 sequences that can be found in the SILVA database and are thus already labeled for the distance computation. We used a threshold of 0.97 (97% similarity) for determining new clusters and a threshold of 0.95 for fuzzy membership. For the choice of each sequence's final cluster, we used different values of α between 0 and 1 with 0.25 increments. The results, presented in Table 3 are compared with those of VSEARCH using identical parameters when applicable.

The program, dataset and corresponding taxonomy are available on `http://projets.isima.fr/sclust/Expe.html`.

3.3 Analysis

We observe that the runtime, while between two to three times longer than VSEARCH's, is still reasonable. The memory usage is slightly higher because more values have to be stored for each sequence/cluster pair. An increase in the importance of the quality (α) results in a decrease in the total number of OTUs and singletons. The proportion of singletons is also reduced. The number of pairs is slightly increased whenever the quality is taken into account but no particular pattern can be discerned. We interpret this as the existence of isolated sequences that initially form singletons but are close enough to be regrouped in small clusters. Unsurprisingly, increasing the importance of the quality increases the average taxonomic distance inside clus-

ters. This increase is not linear and too much emphasis on the quality ($\alpha > 0.75$) drastically increases the distance.

4 Discussion

We observe that the experimental results confirm that adding uncertainty to the clustering helps improve the quality of the output by reducing the number of singletons. Using fuzzy clusters, we are able to extend the clustering threshold to gather additional information on the OTUs's surroundings and use it to quickly assess their quality. This quality can be used together with the distance to choose an OTU for each sequence. The resulting output contains less singletons and misclassifications. Being able to choose the weight of both distance and quality allows for additional tuning.

As previously mentioned, the fuzziness also makes it possible to detect ambiguous sequences and clusters. In our opinion, this is where further work is required. An ambiguous sequence could be arbitrarily assigned to a nearby OTU, become the center of its own OTU or even be considered as an error and deleted but these operations imply such a knowledge of the domain that interactions with the human user become necessary. However, on datasets containing millions of sequences, the number of alerts would render manual treatment impractical or even impossible. Automatizing this treatment would require being able to adapt to the type of data, domain and preferences of the user. We suggest that machine learning techniques be introduced in the process to automatically learn how to handle these ambiguities.

Acknowledgements

This work was supported by the European Union's "*Fonds Européen de Développement Régional (FEDER)*" program and the Auvergne-Rhone-Alpes region.

Author Disclosure Statement

No competing financial interests exist.

References

Davide Albanese, Paolo Fontana, Carlotta De Filippo, et al. MICCA: a complete and accurate software for taxonomic profiling of metagenomic data. *Scientific reports*, 5:9743, 2015.

Violaine Antoine, Benjamin Quost, Marie-Hélène Masson, et al. CECM: constrained evidential c-means algorithm. *Computational Statistics & Data Analysis*, 56(4):894–914, 2012. doi: 10.1016/j.csda.2010.09.021. URL `http://dx.doi.org/10.1016/j.csda.2010.09.021`.

Violaine Antoine, Benjamin Quost, Marie-Hélène Masson, et al. CEVCLUS: evidential clustering with instance-level constraints for relational data. *Soft Comput.*, 18(7):1321–1335, 2014. doi: 10.1007/s00500-013-1146-z. URL `http://dx.doi.org/10.1007/s00500-013-1146-z`.

Wei Chen, Clarence K. Zhang, Yongmei Cheng, et al. A comparison of methods for clustering 16s rRNA sequences into OTUs. *PloS one*, 8(8):e70837, 2013.

Robert C. Edgar. Search and clustering orders of magnitude faster than BLAST. *Bioinformatics*, 26(19):2460–2461, 2010.

Isak Gath and Amir B. Geva. Unsupervised optimal fuzzy clustering. *IEEE Transactions on pattern analysis and machine intelligence*, 11(7):773–780, 1989.

Sarra Ben Hariz, Zied Elouedi, and Khaled Mellouli. Clustering approach using belief function theory. In *International Conference on Artificial Intelligence: Methodology, Systems, and Applications*, pages 162–171. Springer, 2006.

Mylène Hugoni, Najwa Taib, Didier Debroas, et al. Structure of the rare archaeal biosphere and seasonal dynamics of active ecotypes in surface coastal waters. *Proceedings of the National Academy of Sciences*, 110(15):6004–6009, 2013.

Weizhong Li and Adam Godzik. Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics*, 22(13):1658–1659, 2006.

Weizhong Li, Limin Fu, Beifang Niu, et al. Ultrafast clustering algorithms for metagenomic sequence analysis. *Briefings in bioinformatics*, page bbs035, 2012.

Frédéric Mahé, Torbjørn Rognes, Christopher Quince, et al. Swarm: robust and fast clustering method for amplicon-based studies. *PeerJ*, 2:e593, 2014.

Airel Pérez-Suárez, José F. Martínez-Trinidad, Jesús A. Carrasco-Ochoa, et al. OClustR: A new graph-based algorithm for overlapping clustering. *Neurocomputing*, 121:234–247, 2013.

Torbjørn Rognes, Tomáš Flouri, Ben Nichols, et al. VSEARCH: a versatile open source tool for metagenomics. *PeerJ*, 4:e2584, 2016.

Simon Roux, Michaël Faubladier, Antoine Mahul, et al. Metavir: a web server dedicated to virome analysis. *Bioinformatics*, 27(21): 3074–3075, 2011.

Sarah L. Westcott and Patrick D. Schloss. De novo clustering methods outperform reference-based methods for assigning 16s rRNA gene sequences to operational taxonomic units. *PeerJ*, 3:e1487, 2015.

Algorithm 1: DBG Clustering principle

Input : A set of sequences

Output: A set of OTUs to which the sequences are assigned

```
1  Clusters = ∅
2  foreach sequence S do
3      foreach known cluster C do
4          Compute distance(S, C)
5      end
6      if a suitable cluster exists then
7          Assign S to it
8      else
9          Create a new cluster with S as the center
10     end
11 end
12 Return Clusters
```

Algorithm 2: Fuzzy DBG Clustering

Input : A set of sequences

Output: A set of OTUs to which the sequences are assigned

```
1  Clusters = ∅
2  foreach sequence S do
3      foreach known cluster C do
4          Compute distance(S, C)
5          Assign S to C with value f_C(S)
6      end
7      if S has not been sufficiently assigned then
8          Create a new cluster with S as the center
9      end
10 end
11 Return Clusters
```

	Time	Memory	#OTUs	#singletons	#pairs	Distance
Fuzzy ($\alpha = 0$)	15:58	1734180	32597	20966	4618	0.52
Fuzzy ($\alpha = 0.25$)	16:39	1734496	32543	20510	4806	0.56
Fuzzy ($\alpha = 0.5$)	17:28	1735256	32220	20023	4830	0.62
Fuzzy ($\alpha = 0.75$)	16:08	1735056	31567	19347	4754	0.69
Fuzzy ($\alpha = 1$)	17:32	1734180	31309	18276	4826	1.03
VSEARCH	6:13	1332400	34184	22129	4803	0.48

Table 3: Experimental results.

Flash presentations

Easy-HLA web application: new tools for *HLA* genotypes studies

Léo Boussamet[1,2], Estelle Geffard[1,2], Alexandre Walencik[1,2,3], Sophie Limou[1,2,4], Pierre-Antoine Gourraud[1,2], Nicolas Vince[1,2]

1 ATIP-Avenir, Centre de Recherche en Transplantation et Immunologie UMR 1064, INSERM, Université de Nantes, Nantes, France
2 Institut de Transplantation Urologie Néphrologie (ITUN), CHU Nantes, Nantes, France
3 EFS Centre-Pays de la Loire, Nantes, France
4 Ecole Centrale de Nantes, Nantes, France

Corresponding Authors: nicolas.vince@univ-nantes.fr, pierre-antoine.gourraud@univ-nantes.fr

Background: *HLA* genes compose one of the most complex and clinically relevant genetic systems. Its main characteristics are an extreme polymorphism [1] leading to more than 21,500 described alleles and a hereditary transmission by block called haplotypes (one paternal and one maternal). *HLA* compatibility is essential in hematopoietic stem cell transplantation (HSCT) as well as solid organ transplant where donor and recipients must share same *HLA* alleles to be considered compatible. Classically, transplant allocation system still applies donor-recipient *HLA* compatibility mostly at the *HLA* allele level. *HLA* typing quickly evolved with a wide array of techniques and various resolutions. *HLA*-matching remains one of the standard immunological triage tests to determine transplant suitability [2]. Nowadays, advancements in epitopes knowledge have led to the innovative concept of HLA-matching at the epitope level [3]. Indeed, different *HLA* alleles often share common epitopes recognized by specific antibodies.

Aim/Results: We developed Easy-HLA website modules (hla.univ-nantes.fr). These tools allow researchers/clinicians to go further in the understanding of donor/recipient *HLA* compatibilities. First, HLA-Upgrade takes an individual genotype as input, which can be any combination of low/high resolution HLA alleles or even missing ones, and converts it as high resolution outputs based on HLA haplotypes frequencies within a particular population. We provide output genotypes probabilities. Second, HLA-2-Haplo addresses the issue of determining haplotype pairs from genotypes data. It gives in output all possible pairs of haplotypes and their probability of occurrence in a given population. This module also embeds different complementary functions such as HLA-AA, allowing the study of amino acids variation within a sequence, HLA-C expr, predicting HLA-C alleles' expression, HLA-KIRlig predicting the KIR motifs associated with each allele, and HLA-epi translating *HLA* alleles into a combination of epitopes carried by each allele. Finally, HLA-Epi, module sticking to current scientific knowledge regarding epitope-matching was implemented. This module uses the public International Registry of HLA epitopes Epregistry database to compare donors/recipient compatibility in terms of epitopes-matching. In other terms, it will calculate epitopes matches and mismatches between donors and recipient. Every epitope presents in the donor's *HLA* registry but not in the recipient is counted as a mismatch. In a close future, this type of tool should be able to guide clinicians' decisions, especially when no HLA antigen perfect-match donor is found.

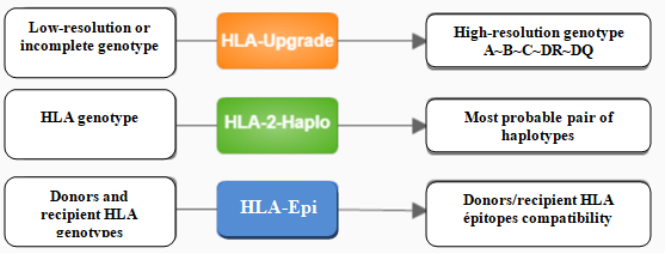


Figure 1Overview of the Easy-HLA complementary modules

Conclusion/perspective: altogether, these different tools enable a broader study of *HLA* genotypes. All gathered on a unique user-friendly open access website, these modules are complementary, and implement the last scientific knowledge dealing with *HLA* compatibilities.

References:

1. Robinson,J. et al. (2015) The IPD and IMGT/HLA database: allele variant databases. Nucleic Acids Res.
2. Zachary,A.A. and Leffell,M.S. (2016) HLA Mismatching Strategies for Solid Organ Transplantation – A Balancing Act. Front. Immunol.
3. Lachmann,N. *et al.* (2017) Donor-Recipient Matching Based on Predicted Indirectly Recognizable HLA Epitopes Independently Predicts the Incidence of De Novo Donor-Specific HLA Antibodies Following Renal Transplantation. *Am. J. Transplant.*

Key words: HLA epitope, Easy-HLA, HLA genotype, solid organ transplantation, HSCT.

Évaluation d'outils d'identification et de quantification des transcrits alternatifs à partir de données de séquençage longue lecture Nanopore

Bérengère Laffay[1,2], Corinne Blugeon[1], Fanny Coulpier[1], Stéphane Le Crom[1,3], Sophie Lemoine[1] and Laurent Jourdren[1]

1. Institut de biologie de l'École normale supérieure (IBENS), École normale supérieure, CNRS, INSERM, PSL Université Paris 75005 Paris, France.

2. Master Bioinformatique, Normandie Université, UNIROUEN.

3. Sorbonne Université, Univ Antilles, Univ Nice Sophia Antipolis, CNRS, Evolution Paris Seine-Institut de Biologie Paris Seine (EPS-IBPS), F-75005 Paris, France.

Corresponding author: laffay@biologie.ens.fr

Keywords: MinION, Oxford Nanopore, pipeline, analysis, alternative splicing

La technologie Oxford Nanopore permet de produire des lectures de plusieurs kilobases. En génomique fonctionnelle, elle offre un accès direct aux transcrits (sans étape d'assemblage) et ainsi la possibilité de caractériser et quantifier les séquences ARN à l'échelle des transcrits alternatifs (ou isoformes). La stratégie qui a été adoptée pour identifier les transcrits alternatifs est de construire un transcriptome en regroupant les lectures suivant leurs jonctions d'épissages. Nous avons retenu deux outils développés à cet effet disponibles dans la communauté : Flair [1] et Pinfish [2].

Flair (Full-Length Alternative Isoform analysis of RNA) est le pipeline développé en Python 2.7 par le laboratoire dirigé par Angela Brooks à l'Université de Californie [3] pour proposer le traitement des lectures longues. Flair permet de procéder de l'alignement des lectures issues des technologies de séquençage Oxford Nanopore ou PacBio jusqu'à l'analyse de l'expression différentielle en passant par la correction des sites d'épissages via les sites d'épissages de l'annotation de référence et si souhaité, via les lectures courtes Illumina. Ces transcrits sont définis à partir des combinaisons de sites d'épissages unique. Des modules sont prévus dans le pipeline pour procéder au comptage par transcrits alternatifs avec Salmon[4] et à l'analyse différentielle.

Il a notamment été utilisé par le nanopore-ngs-consortium pour évaluer la pertinence des séquençages d'ADNc et d'ARN natifs chez l'Homme[5]. Peu après, Oxford Nanopore a développé Pinfish[6] en langage Go. Ce pipeline inspiré par le pipeline Mandolarion[7], est destiné à générer des annotations à partir de lectures longues. De la même manière que Flair, les transcrits qu'il fournit sont caractéristiques des suites de sites d'épissages observées à la différence près que les transcrits alternatifs sont construits à partir de la médiane des bornes des exons de chaque groupe. Une étape de correction avec Racon[8] des transcrits alternatifs identifiés est comprise dans son workflow snakemake.

Ces deux outils proposent plus ou moins de souplesse sur les paramètres de caractérisation des transcrits. Ce poster a pour objet de présenter les résultats de leurs performances sur des données RNA-Seq séquencées chez la souris avec la chimie 1D d'Oxford Nanopore, entre complexité d'installation, temps de traitement et comparaison des ensembles de transcrits retenus par chacun.

[1] https://github.com/BrooksLabUCSC/flair

[2] https://github.com/nanoporetech/pinfish

[3] Full-length transcript characterization of *SF3B1* mutation in chronic lymphocytic leukemia reveals downregulation of retained introns ; Alison D Tang, Cameron M Soulette, Marijke J van Baren, Kevyn Hart, Eva Hrabeta-Robinson, Catherine J Wu, Angela N Brooks ; bioRxiv 410183; doi: https://doi.org/10.1101/410183

[4] Patro, R., Duggal, G., Love, M.I., Irizarry, R.A. & Kingsford, C. Salmon provides fast and bias-aware quantification of transcript expression. Nat. Methods **14**, 417–419 (2017).

[5] Nanopore native RNA sequencing of a human poly(A) transcriptome ; Rachael E Workman, Alison Tang, Paul S. Tang, Miten Jain, John R Tyson, Philip C Zuzarte, Timothy Gilpatrick, Roham Razaghi, Joshua Quick, Norah Sadowski, Nadine Holmes, Jaqueline Goes de Jesus, Karen Jones, Terrance P Snutch, Nicholas James Loman, Benedict Paten, Matthew W Loose, Jared T Simpson, Hugh E. Olsen, Angela N Brooks, Mark Akeson, Winston Timp; bioRxiv 459529; doi: https://doi.org/10.1101/459529

[6] https://community.nanoporetech.com/knowledge/bioinformatics/using-pinfish-for-gene-tra/tutorial

[7] R2C2: Improving nanopore read accuracy enables the sequencing of highly-multiplexed full-length single-cell cDNA ; Roger Volden, Theron Palmer, Ashley Byrne, Charles Cole, Robert J. Schmitz, Richard Edward Green, Christopher Vollmers; bioRxiv 338020; doi: https://doi.org/10.1101/338020

[8] Li H, Durbin R. 2009. Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics 25:1754–1760. doi:10.1093/bioinformatics/btp324. https://github.com/isovic/racon

From genomics to metagenomics: benchmark of variation graphs

Kévin DA SILVA[1,2], Nicolas PONS[1], Magali BERLAND[1], Florian PLAZA-OÑATE[1], Mathieu ALMEIDA[1], Pierre PETERLONGO[2]

[1] Univ. Rennes, 2 rue du Thabor, 35042, Rennes, France
[2] MetaGenoPolis, INRA, Université Paris-Saclay, 78350, Jouy-en-Josas, France

Corresponding Author: kevin.da-silva@inria.fr

In the metagenomics field, the classical approach for quantitative analysis of sequencing data consists into aligning sequence reads to a non redundant reference gene catalogue that represents a specific ecosystem [1]. However this approach lacks flexibility and exhaustiveness as it uses a fixed catalogue built upon a limited number of samples. To overcome those biases, the reads could be aligned to a more informative reference structure covering the variants encountered in the population and also more complete with a full genome catalogue thus giving access to the genomic structure (including operonic and intergenic regions). Recently, the pangenome concept has been increasingly used as it opens new ways to investigate multiple genomes of close individuals, as for characterizing the different strains of a species. Associations between strain variants and phenotype are then of great interest for diagnostic and therapeutic strategies.

Erik Garrison et al. have developed "*vg*", a toolkit for creating variation graphs, bidirected DNA sequence graphs that represents multiple genomes, including their genetic variation [2]. This structure, together with the set of paths drawn by the ordered consecutive nodes of the graph containing single strand DNA sub-sequences, allows the possible sequences from a population to be accessed, and thus provides a way to represent the pan-genome of a species. With a perspective towards metagenomics, we foresee *vg* as a tool enabling to build a catalogue of pangenomes from metagenomic samples. Our goal is to identify and characterize each species by using contigs binning and path optimization, and represent each species as a single variation graph. Additionally, we assist to a huge increase of available genomes issued from cultivated isolates [3] or metagenomics assemblies [4] which need to be addressed. The variation graphs could then be a mean of integrating all the current and future information.

As a proof of concept, we started back to a genomic level using *E. coli* for its variability between strains depending of the pathogenicity. Complete genomes of six strains, pathogenic and non-pathogenic, were selected to build a variation graph. Among them, the strain O104:H4 was selected as it has been studied during the outbreak of shiga-toxigenic *E. coli* (STEC), which struck Germany in May-June 2011 [5]. The first step was to benchmark *vg* to have a global view of the computation time to build a graph considering different inputs: complete genomes or chopped parts of the complete genomes thus simulating contigs of various lengths. Secondly, reads without errors were simulated for each strain and mapped back on the variation graph in order to check the validity of the graph. This was accomplished through read counting on paths of the graph, each path corresponding to a strain or contigs of the strains, and allowing the identification of the strain which has generated the reads.

We will present the results using the same methodology on real data, showing that reads from the German outbreak study can be used to check the STEC-positive and -negative samples using the variation graph. We will also discuss the scalability of this approach on a metagenomic level and identify the possible issues or biases on a mock data composed of almost a hundred species.

References

[1] Li J, et al. An integrated catalog of reference genes in the human gut microbiome. Nat. Biotechnol. 32, 834–841, 2014.
[2] Garrison E, et al. Variation graph toolkit improves read mapping by representing genetic variation in the reference. Nat Biotechnol. 36(9):875–9, 2018.
[3] Zou, Y, et al. 1,520 reference genomes from cultivated human gut bacteria enable functional microbiome analyses. Nat. Biotechnol. 37, 179, 2019
[4] Pasolli E, et al. Extensive unexplored human microbiome diversity revealed by over 150,000 genomes from metagenomes spanning age, geography, and lifestyle. Cell176, 649–662, 2019.
[5] Loman NJ, et al. A culture-independent sequence-based metagenomics approach to the investigation of an outbreak of Shiga-toxigenic Escherichia coli O104:H4. JAMA 309:1502–1510, 2013.

Genomic evolution of contralateral breast cancer revealed from whole exome sequencing

Zakia TARIQ[1], Florian BONIN[1], Claire FAYARD[1], Ivan BIECHE[1], Virginie RAYNAL[2], Sylvain BAULANDE[2], Rosette LIDEREAU[1] and Keltouma DRIOUCH[1]

[1] Service de génétique, Institut Curie, 26 rue d'Ulm 75248, Paris, France

[2] Next generation sequencing platform, ICGex, Institut Curie, Paris, France

Corresponding Author: zakia.tariq@curie.fr

1. Background

Approximately 2-10% of women with breast cancer developed a tumor in the contralateral breast, associated with worse prognosis than unilateral cancer. Despite the advance in high throughput sequencing, evolution processes underlying these bilateral breast cancers remain poorly understood. Currently, both tumors are considered as independent primary cancers and treated accordingly.

This project aims at testing whether these tumors could be clonally related; one being the metastasis of the other. To this end, we studied the mutational profiles of a series of contralateral breast tumors obtained though whole exome sequencing.

2. Methods

Matched bilateral breast tumors and germline DNAs were obtained from 20 patients, divided in 2 groups: 9 with synchronous bilateral breast tumors and 11 primary tumors with paired contralateral tumor that underwent first line adjuvant treatments. On Illumina platform, paired-end 100x100, Exome-Seq was performed and three variant callers were used in parallel to predicted variations (Mutect1, HaplotypeCaller and UnifiedGenotyper). After Annovar annotation, variants with low 1000Genome frequency and absent in germline samples, were validated by IGV visualization and then considered as somatic variants. To characterize the recurrent drivers of breast cancer progression, we first annotated our variants on the basis of referenced cancer genes (Cancer Gene Census database and bibliography) and then identified new drivers thought Mutational Significance in Cancer tool (MuSiC).

Copy number alterations were determined by Facets and samples purity and clonality were established by ABSOLUTE tool. Finally, PyClone was used to predict cancer evolution process of each patient.

3. Results

Most of patients exhibit different copy number profiles and cancer driver genes are mutated independently in each tumor pairs. But for around 50% of patients, we observed same copy number patterns and hotspot mutations in both tumors. Phylogenetic analyses of these contralateral breast cancer revealed a clonal evolution of these tumors, in both synchronous and metachronous groups. The difference of trunk part size between metachronous and synchronous phylogenetic trees, revealed an earlier metastatic divergence in synchronous contralateral breast tumors.

This study highlights the utility of discriminated independent primary cancer from breast metastases, which seems to be underestimated. Our findings might be important for the clinical management of breast cancer patients since metastases are more aggressive tumors than primary cancers and require appropriate systemic treatments.

Inter-individual variability in healthy human cytokine responses

Violaine SAINT-ANDRE[1,2], Vincent ROUILLY[3], Bruno CHARBIT[4], Anne BITON[1], Matthew ALBERT[3], Lluis QUINTANA-MURCI[5], Darragh DUFFY[2] and the *Milieu Intérieur* Consortium

[1] Institut Pasteur, Bioinformatics and Biostatistics Hub, C3BI, USR756 IP CNRS, Paris, France

[2] Institut Pasteur, Immunobiology of Dendritic Cell Unit, Inserm U1223, Paris, France

[3] Department of Cancer Immunology, Genentech, South San Francisco, USA

[4] Institut Pasteur, Center for Translational Research, Paris, France

[5] Unit of Human Evolutionary Genetics, Institut Pasteur, Paris, France

Corresponding Author: violaine.saint-andre@pasteur.fr

While response to stress and infection are known to vary across individuals, medical practices and public health policies remain based on a single model of patient care and drug development. The *"Milieu Intérieur"* project was developed to help the transition from these practices towards precision medicine. This project primarily aims to define the boundaries of a "healthy" immune response and to assess which are the genetic, epigenetic and environmental factors that influence its variation between individuals. Within the frame of this project, we have analyzed immune response variability quantifying 13 selected cytokines in 12 immune stimulation conditions for a cohort of 1,000 well-define healthy donors. Integration of socio-demographic, clinical, nutritional and environmental factors together with proteomic and transcriptomic data from these individuals points to the specific factors that influence their expression in each immune stimulation condition, and may help us to predict how they contribute to susceptibility to infection, therapeutic treatment or vaccine response.

Keyworks: data integration, variable selection, linear models, cohort study, immune response

Panache: a visualization tool for the exploration of plant pangenomes

Éloi DURANT[1,2,3,4], François SABOT[1,4], Matthieu CONTE[3] and Mathieu ROUARD[2,4]

[1] DIADE University of Montpellier - IRD, 911 avenue Agropolis, 34934, Montpellier, France
[2] Bioversity International, Parc Scientifique Agropolis II, 34397, Montpellier, France
[3] Syngenta France, 12 Chemin de l'Hobit, 31790, Saint-Sauveur, France
[4] South Green Bioinformatics Platform, Bioversity – CIRAD – INRA – IRD, Montpellier, France

Corresponding Author: eloidurant@gmail.com

Low-cost high-throughput sequencing technologies enabled the production of tremendous amount of genomic data, including multiple genomic sequences for a single species. Comparisons of such sequences showed that there are structural variations even between individuals from the same species, like Copy Number Variations (CNV) and Presence/Absence Variations (PAV) [1]. Thus, a single reference genome is insufficient to grasp all these variations.

Pangenomics is an integrative approach which aims to the assessment of such genomic variations and more within a group of closely related individuals. It can take slightly divergent meanings depending on the studies, mainly based on either a functional or a structural approach: its definition can be focused on the whole repertoire of genes within a group or can include blocks of genomic sequences more or less shared between species [2]. Although it has already often been applied to bacteria, its use with plants is still sparse but long-reads technologies might generalize it significantly [3]. Pangenomics faces many computational challenges, partly because of its relative novelty and partly because of its similarities with 'Big-Data'. New tools are needed for assessing the problems of its construction, storage and analysis, but also visualization [4].

Existing and usable tools for bacterial pangenomics (*Anvi'o*, *PanACEA* …) yet only focus on genes and do not enable any structural exploration of plant pangenomes. Nowadays, two main approaches of visualization for structural pangenomics are studied. The first one is *graph*-based, with genome sequences sliced into pieces that can be shared between genomes or specific to one. Those pieces are represented as nodes in a network, connected together depending on their order in the original genomes. Therefore there are as many paths connecting the nodes as genomes used to build the pangenome. *PPanGGOLiN* and the *Augmented Graph Viewer* (*agv*) are both examples of tools being developed to work with such a representation. The second approach is a *linear* one, adapted from existing genome browsers. The information of presence and absence of genome parts is displayed along a pangenomic reference.

We would like to introduce here our own visualization tool, based on that second representation: the *PANgenome Analyzer with CHromosomal Exploration* (*Panache*). *Panache* is a web-based application which enables its users to explore a pangenomic reference divided in multiple *panchromosomes*. For now it allows a quick identification of genomic blocks belonging to either the *core* or *dispensable* genomes along with the corresponding Presence/Absence Matrix, and navigation within and between panchromosomes. The prototype presented today is the base for further work that aims to add functionalities like the identification of repeated blocks within the pangenome, filtering and ordering of blocks and constitutive genomes, variant calling, retrieval of the original order of blocks for a certain genome...

References

1. Nathan M Springer et al., Maize Inbreds Exhibit High Levels of Copy Number Variation (CNV) and Presence/Absence Variations (PAV) in Genome Content, *PLOS Genetics*, 5(11), 2009.
2. Christine Tranchant-Dubreuil, Mathieu Rouard and François Sabot, Plant Pangenome: Impacts On Phenotypes And Evolution, *Annuals Plant Reviews Online*, In press, 2019.
3. Agnieszka A Golicz, Jacqueline Batley and David Edwards, Towards plant pangenomics. *Plant Biotechnology Journal*, 14(4):1099-1105, 2016.
4. The Computational Pan-Genomics Consortium, Computational pan-genomics: status, promises and challenges, *Briefings in Bioinformatics*, 19(1):118-135, 2018

scViz: a Rshiny app to easily explore scRNAseq data

Wilfrid RICHER[1,3], Jimena TOSSELO[1,3], Solène BROHARD[2,3], Joshua WATERFALL[2,3], Eliane PIAGGIO[1,3]

[1] Unité INSERM U932, Institut Curie, Paris, France
[2] Unité INSERM U830, Institut Curie, Paris, France
[3] Département de Recherche Translationnelle, Institut Curie, Paris, France

Corresponding Author: wilfrid.richer@curie.fr

Abstract

We present a Rshiny application whose function is to offer biologists the capability to easily explore and visualize single cell RNAseq data integrating multiple analysis tools.

This application requires a .Rds or a .Rdata formatted input file providing a S4 class object returned by the Seurat pipeline [1,2] (containing the pre-computed data: the expression matrix, the metadata specifying features such as cluster identities, patient annotations and experimental batches, as well as the dimensionality reduction matrix). This file can contain the analysis of a single sample or the aggregate of multiple samples.

The Rshiny application currently offers the possibility to visualize and explore the data using multiple approaches: 1) Visualization of the evolution of the clusters according to the resolution to aid understanding of inter-cluster relatedness and robustness [3]; 2) Dimensionality reduction such as t-SNE [4] or UMAP [5] with cluster assignment for cells which takes into account the resolution chosen by the user; 3) Barplot to visualize the composition of the clusters as a percentage of cells according to the metadata present in the loaded object (e.g. cell types, patients, batches, resolutions); 4) Dimensionality reduction of the data to visualize the expression of a gene or a list of genes within the data; 5) Violinplot comparing the expression of a gene or a set of genes within the different clusters of the resolution chosen by the user.

For the convenience of the user, the scViz tool offers several features including the quick search of a gene by autocompletion, personalization of the figures by allowing the modification of the axes and the size of the component elements, and the removal of specific clusters. The generated figures can furthermore be exported in high quality image files (600dpi).

The scViz application was tested on a 10X scRNAseq dataset (V2 chemistry on 3' and 5' UTR) consisting of approximately 50,000 CD4+ T cells derived from blood, lymph nodes and tumors of 5 lung cancer patients. ScViz provided an aid to visualize the diversity of CD4+ T cell populations (both conventional and regulatory T cells). This made it possible to confirm subpopulations characterized by the expression of specific genes and to illustrate the composition of the different populations according to their original tissues.

This graphical interface is under active development following the feedback of users, making it more intuitive for a user to perform exploratory analyses of diverse single cell RNAseq datasets. This Rshiny app is available in GitHub (https://github.com/wilfridricher/scViz) for the analysis of single cell RNAseq data. It runs for version 2 and for version 3 of Seurat pipeline.

Acknowledgements

We acknowledge all the users of scViz whose helped in development and testing of this Rshiny application.

References

1. Butler A et al. Integrating single-cell transcriptomic data across different conditions, technologies, and species. Nat Biotechnol. 2018 Jun;36(5):411-420. doi: 10.1038/nbt.4096.

2. Tim S et al. Comprehensive integration of single cell data. bioRxiv 460147; doi: https://doi.org/10.1101/460147.
3. Zappia L, Oshlack A. Clustering trees: a visualization for evaluating clusterings at multiple resolutions. Gigascience. 2018;7. doi: gigascience/giy083.
4. McInnes, L, Healy, J, Melville J. UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction, ArXiv e-prints 1802.03426, 2018
5. Van der Maaten, L.J.P.; Hinton, G.E. Visualizing Data Using t-SNE. Journal of Machine Learning Research. 2008 Nov(9): 2579–2605

Using metabolomic data to predict Maize yield

Sylvain Prigent[1], Olivier Fernandez[1,2], Stéphane Bernillon[1,2], Pierre Petriacq[1,2], Annick Moing[1,2], Thierry Berton[1,2], Llorenç Cabrera-Bosquet[3], Émilie Millet[3], Claude Welcker[3], François Tardieu[3] and Yves Gibon[1,2]

[1] UMR 1332 BFP, INRA, Univ. Bordeaux, F33883 Villenave d'Ornon, France
[2] Plateforme Métabolome du Centre de Génomique Fonctionnelle Bordeaux, France
[3] LEPSE, INRA, Univ. Montpellier, 34060 Montpellier, France

Corresponding author: sylvain.prigent@inra.fr

The metabolome is often seen as the ultimate phenotype, resulting from all biochemical processes taking place in an organism. Such a phenotype could thus contribute to the prediction of yield [1], and give insight into how genetic differences in yield arise. In this study, we describe how the analysis of metabolites present in maize leaves grown in a greenhouse under two conditions (drought stress or control) served to estimate grain yield of plants grown in different fields.

Targeted microplate metabolite analysis and liquid-chromatography coupled with mass-spectrometry (LC-MS) based profiling were performed on 238 genotypes of maize plants grown at the PhenoArch platform [2], under both well-watered and drought conditions, yielding data for 11 major metabolites and 1,415 metabolite signatures, respectively. Grain yield were collected from the same genotypes in 15 different fields across Europe [3]. Multilinear modelling was then used to link metabolic patterns to plant performance. Given the high number of variables compared to the number of individuals, to limit risks of overfitting, systematic cross-validation was performed.

The proposed models were able to perform good predictions with, depending on the different field conditions tested, correlations ranging from +0.5 to +0.65. Interestingly, the best yield predictions were obtained for well-watered fields based on metabolomic data gathered from stressed plants. Currently, the best models still need many variables to perform the predictions, making their interpretation complicated. In a next step, a tentative manual annotation of the MS-based metabolic variables that are clearly linked to yield performance in specified scenarios will be performed.

Acknowledgements

We thank MetaboHUB (ANR-11-INBS-0010), PHENOME (ANR-11-INBS-0012) and AMAIZING (ANR-10-BTBR-01) projects for financing, PHENOARCH team for growing the plants in controlled conditions and DROPS (EU FP7–244374) for field phenotyping.

References

[1] Shizhong Xu, Yang Xu, Liang Gong, and Qifa Zhang. Metabolomic prediction of yield in hybrid rice. *The Plant Journal*, 88(2):219–227, 2016.

[2] Llorenç Cabrera-Bosquet, Christian Fournier, Nicolas Brichet, Claude Welcker, Benoît Suard, and François Tardieu. High-throughput estimation of incident light, light interception and radiation-use efficiency of thousands of plants in a phenotyping platform. *New Phytologist*, 212(1):269–281.

[3] Emilie J. Millet, Claude Welcker, Willem Kruijer, Sandra Negro, Aude Coupel-Ledru, Stéphane D. Nicolas, Jacques Laborde, Cyril Bauland, Sebastien Praud, Nicolas Ranc, Thomas Presterl, Roberto Tuberosa, Zoltan Bedo, Xavier Draye, Björn Usadel, Alain Charcosset, Fred Van Eeuwijk, and François Tardieu. Genome-wide analysis of yield in europe: Allelic effects vary with drought and heat scenarios. *Plant Physiology*, 172(2):749–764, 2016.

Software demonstrations

A precision medicine application: personalized contextualization of patients after kidney transplantation.

Estelle GEFFARD[1,2], Pauline SCHERDEL[1,2], Sophie LIMOU[1,2,3], Sophie BROUARD[1,2], Magali GIRAL[1,2], Nicolas VINCE[1,2], Pierre-Antoine GOURRAUD[1,2]

[1] Centre de Recherche en Transplantation et Immunologie UMR 1064, INSERM, Université de Nantes, Nantes, France

[2] Institut de Transplantation Urologie Néphrologie (ITUN), CHU Nantes, Nantes, France

[3] Ecole Centrale de Nantes, Nantes, France

Corresponding Author: nicolas.vince@univ-nantes.fr, pierre-antoine.gourraud@univ-nantes.fr

Around 14% of adults suffer from chronic kidney disease (CKD). Among those, most patients with advanced illnesses evolve to end-stage disease, and become candidate for kidney transplantation. We developed KiTapp (Kidney Transplantation application) a precision medicine application for kidney transplantation patients in order to monitor individual trajectories of post-transplant outcomes using data from the DIVAT (Données Informatisées et VAlidées en Transplantation) study. Together, information from more than 3,000 patients with renal transplantation, including clinical and immunological items, were collected since 2008. We present here two personalized contextualization algorithms: 1) a populational contextualization where we compare data trajectory of a given patient to a sub-population with similar characteristics selecting by filters or nearest neighbor approaches, and 2) a referential contextualization where we compare data trajectory of a given patient to extreme groups (defined by clinicians) such as acute graft rejection, humoral rejection, cellular rejection or tolerance. The comparative properties of these algorithms are individually determined at reference group level. An example is the normality assessment of the 50-, 100-, 500-, 1000-, or 1500- nearest neighbors for a given kidney transplantation patient with a 150mol/L creatinine level at 1 year post-transplantation; we find him at the 65e, 57e, 67e, 67e, or 66e percentiles, respectively. We developed a R Shiny prototype (Figure). This precision medicine application therefore facilitates access to large amount of data and allows their visualization and comparison in order to optimize medical care and guide clinical decision. Finally, we ambition to extend our algorithms to various chronic medical conditions and settings to improve patient's care.

Figure : Application prototype of precision medicine

Keywords

Kidney transplantation, precision medicine, contextualization algorithms, KiTapp, DIVAT

Allogenomics – pipeline: prediction of the immune response from genetic variants during transplantation

Robert CLERC[1,2], Hugues RICHARD[2] and Laurent MESNARD[1,3]

[1] Institut des Sciences des données et du Calcul, Sorbonne Université, 4 place Jussieu 75005 Paris

[2] LCQB UMR7238 CNRS, Institut de Biologie Paris Seine, Sorbonne Université, 4 place Jussieu 75005 Paris

[3] Inserm UMR-S 1155, Hôpital Tenon, Sorbonne Université, 4 rue de la chine 75020 Paris

Corresponding Author: robert.clerc@sorbonne-universite.fr

The immune response of a genome to a host cell is a question that concerns multiple areas of research. It can be used for the prediction of immune response in the case of a transplant organ where the immune system of the recipient is activated [1], The same apply to predict the sensibility to immunotherapy [2], as in others research projects requiring the study of the adaptive immune response.

We propose here a pipeline, which can be used in a reproducible way for research projects requiring to predict binding affinity of immunogenic peptides using genotype from exome sequencing data. From a given vcf file containing at least two individuals that has been annotated with Variant Effect Predictor (VEP) from Ensembl [3], we developed the allogenomics pipeline (Figure 1) which enable:

1 – Computation of the Allogenomics Mismatch Score (AMS) by extracting and annotating non-synonymous SNPs along the coding fraction of the genome *mismatched* between the two individuals. The tool then builds the set of peptides around those mismatch positions.

2 – Affinity prediction: The affinity of all peptides found in mismatch is then predicted with a third-party software. We use NetMHCPan in our case [4]

3 – A2MS: Prepare a list of candidate peptides annotated according to our assumption.

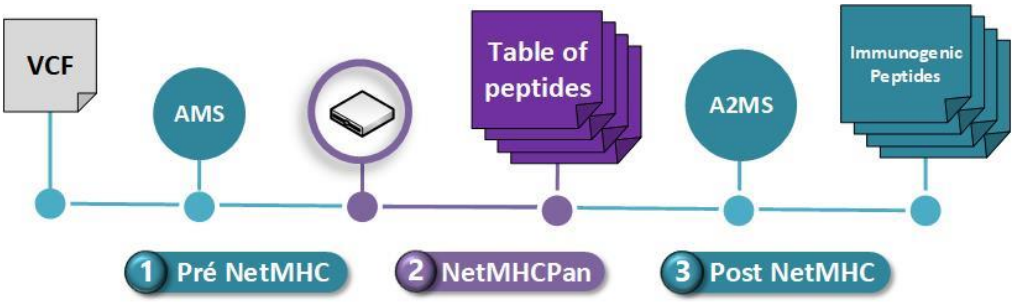


Figure 1: The Allogenomics Pipeline leading to the AMS and the A2MS. Blue: pipeline and files developed by Allogenomics team. Purple: third party Software [4].

References

[1] Mesnard, et al., Exome Sequencing and Prediction of Long-Term Kidney Allograft Function. PLoS Comput Biol. 2016 Sep 29;12(9):e1005088

[2] Heindl et al, Microenvironmental niche divergence shapes BRCA1-dysregulated ovarian cancer morphological plasticity. Nature Communication 2018 9:3917

[3] Daniel R. and all Ensembl 2018. PubMed PMID: 29155950.

[4] NetMHCpan-4.0: Improved Peptide–MHC Class I Interaction Predictions Integrating Eluted Ligand and Peptide Binding Affinity Data, Vanessa Jurtz, Sinu Paul, Massimo Andreatta, Paolo Marcatili, Bjoern Peters and Morten Nielsen, *The Journal of Immunology* (2017) ji1700893; DOI: 10.4049/jimmunol.1700893

EasyMatch-R: a web application to facilitate donor query in Hematopoietic Stem Cell Transplantation (HSCT)

Estelle GEFFARD[1,2], Alexandre Walencik[1,2,3], Sophie LIMOU[1,2,4], Anne CESBRON [3], Nicolas VINCE[1,2], Pierre-Antoine GOURRAUD[1,2]

[1] ATIP-Avenir, Centre de Recherche en Transplantation et Immunologie UMR 1064, INSERM, Université de Nantes, Nantes, France

[2] Institut de Transplantation Urologie Néphrologie (ITUN), CHU Nantes, Nantes, France

[3] EFS Centre-Pays de la Loire, Nantes, France

[4] Ecole Centrale de Nantes, Nantes, France

Corresponding Author: nicolas.vince@univ-nantes.fr, pierre-antoine.gourraud@univ-nantes.fr

HLA genes compose one of the most complex and clinically relevant genetic system. HLA matching is essential in hematopoietic stem cell transplantation (HSCT) to determine compatibility between one patient and potential donors. The extreme HLA genes polymorphism (21,500 described alleles so far) and the constant refinement of HLA typing techniques complicate compatible individuals lookup among 30 millions worldwide recorded volunteer bone marrow donors. We have developed Easy-HLA, a suite of web tools to simplify the study of HLA in an individual or cohort (hla.univ-nantes.fr). Easy-HLA computes a statistical method of HLA haplotypes inference based on their frequencies in a population. Easy-HLA works with a critical database of more than 600,000 haplotypes representative of the world population and their frequencies among 5 large populations. These haplotype frequencies are derived from 6.59 million donors genotypes of the National Marrow Donor Program[1]. The haplotypes and their frequencies were calculated with a maximization estimation algorithm from the HLA genotypes. Easy-HLA's imputation algorithm is based on these haplotypic frequencies. When HLA genotypes are specified with ambiguities and/or incomplete (noted XX on equation below), several pairs of alternative haplotypes (p1-p4) can be deduced. Haplotypes not present in the reference database are removed from the list of haplotypes. From an incomplete HLA genotype (G), the Easy-HLA algorithm produces all possible pairs of haplotypes and then calculates their corresponding probability.

Equation: enumeration of haplotype pairs of an incomplete genotype with a missing locus

$$G(\mathrm{A\,a} \sim \mathrm{XX} \sim \mathrm{Cc}) \begin{cases} \mathrm{p1}\ (\mathrm{A} \sim \mathrm{B} \sim \mathrm{C}, \mathrm{a} \sim \mathrm{b} \sim \mathrm{c}) \\ \mathrm{p2}\ (\mathrm{A} \sim \mathrm{b} \sim \mathrm{C}, \mathrm{a} \sim \mathrm{B} \sim \mathrm{c}) \\ \mathrm{p3}\ (\mathrm{A} \sim \beta \sim \mathrm{c}, \mathrm{a} \sim \mathrm{b} \sim \mathrm{C}) \\ \mathrm{p4}\ (\mathrm{A} \sim \beta \sim \mathrm{c}, \mathrm{a} \sim \beta \sim \mathrm{C}) \end{cases}$$

EasyMatch-R, a component of Easy-HLA, calculates the probability of finding an HLA matched HSCT donor in a given population and recommends parsimonious complementary HLA typing strategy required before requesting blood sample. One major challenge in practice for clinical laboratories is to trust and integrate such "statistical decision support" programs. We present here a comparative analysis of searches and strategies performed on 202 patients with HSCT indication from Nantes. When comparing number of expected donors obtained from EasyMatch-R and volunteer donors registered in the BMDW book, we found a significant positive correlation ($r=0.80$, $p=3.9\times10^{-46}$). We retrospectively compared the impact of the recommendation algorithm and the number of additional typings requested by the lab at the time of the donor search. When considering only individuals with more than 10% chance of being a match with the input HLA type, EasyMatch-R recommends 134 different typings compared to 249 requested for the most likely potential donor from the BMDW book. Overall, this shows that EasyMatch-R facilitates the search of a donor by providing a statistically quantified argument supporting early adoption of alternative options when the BMDW book is not favorable. It improves the effectiveness and diminishes the cost related to additional HLA typing.

References

1. Gragert, L., Madbouly, A., Freeman, J. & Maiers, M. Six-locus high resolution HLA haplotype frequencies derived from mixed-resolution DNA typing for the entire US donor registry. *Hum. Immunol.* **74**, 1313–1320 (2013).

Keywords

Easy-HLA, HLA, HSCT, donor compatibility, blood cancer, personalized medicine, hematopoietic stem cell transplantation

INEX-MED: a Knowledge Graph to explore and link heterogeneous bio-medical data

Maxime Folschette[1,2,⋆], Kirsley Chennen[1,3,4,⋆], Alban Gaignard[5], Richard Redon[5], Hala Skaf-Molli[2], Olivier Poch[3], Jocelyn Laporte[4], Julie Thompson[3] and the INEX-MED consortium

[1] CNRS UMS 3601, Institut Français de Bioinformatique, France
[2] LS2N, Laboratoire des Sciences du Numérique de Nantes, CNRS UMR 6004, Université de Nantes, France
[3] CSTB - iCUBE, CNRS UMR 7357, Faculté de Médecine, Strasbourg, France
[4] Institut de Génétique et de Biologie Moléculaire et Cellulaire (IGBMC), INSERM U1258, CNRS UMR7104, Illkirch, France
[5] Institut du Thorax, Inserm UMR 1087, CNRS UMR 6291, Université de Nantes, France

Corresponding author: maxime.folschette@ls2n.fr, kchennen@unistra.fr, alban.gaignard@univ-nantes.fr

1 Context and Motivations

Health and life sciences nowadays face the massive availability of diverse biomedical data. Providing a unified and coherent access to these large-scale multi-source data is a major challenge. The INEX-MED project aims at gathering and representing multi-disciplinary, multi-source and multi-modal data (clinical, genetic, imaging) as a Knowledge Graph, in the domain of intracranial aneurysms and congenital myopathies. The aim is to apply machine learning in order to make predictions, highlight new diagnosis variables and stratify patient populations. To this end, Linked Data principles are applied in order to integrate the data despite their initial heterogeneity, with the objective of ensuring "FAIR" data principles (Findability, Accessibility, Interoperability, Reusability) [1].

2 Knowledge Graph Creation

The acquired diverse data, come in different formats. For instance, clinical data are represented with tables (CSV format), while genetic (exome) data rely on the VCF format. Imaging data come in specific formats and are automatically processed to extract quantitative markers. All data are then represented in a directed labelled graph (RDF format). Many existing domain-specific ontologies were explored to find relevant concepts and relations.

This knowledge graph can then be queried with SPARQL, a graph-pattern based query language designed to select nodes or edges, or assemble sub-graphs. The main advantage of this approach is that such queries can relate to all parts of the data at once (clinical, genetic, imaging) without the need for explicit joins. Our implementation thus offers convenient multi-source data access: it is now possible to return, for instance, clinical features (phenotype, diagnosis, ...) of individuals having a genetic variant on a specific set of genes, with a single SPARQL query. In addition, federated queries allow to pull data from external sources (for instance, Orphanet and Uniprot) and thus dynamically enrich the knowledge graph.

3 Demonstration scenario

We propose to showcase our prototype in the form of a web application dedicated to biologists and a Jupyter Notebook dedicated to bioinformaticians.

We will demonstrate how these interfaces directly interact with the knowledge graph in the form of template SPARQL queries and how they can be used to answer biological questions. We will show results as graphical plots, for monitoring purposes, or tables, for further bio-statistics analysis or machine learning based predictive modelling.

References

[1] Mark D Wilkinson, Michel Dumontier, IJsbrand Jan Aalbersberg, Gabrielle Appleton, Myles Axton, Arie Baak, Niklas Blomberg, Jan-Willem Boiten, Luiz Bonino da Silva Santos, Philip E Bourne, et al. The FAIR Guiding Principles for scientific data management and stewardship. *Scientific data*, 3, 2016.

⋆. These authors contributed equally to this work

Leaves : Application d'aide à l'interprétation de variants

Vivien DESHAIES[1], Mathieu BARTHELEMY[1], Alban LERMINE[1, 2]

[1] MOABI, 33 boulevard Picpus, 75012, Paris, France
[2] SeqOIA-IT, 33 boulevard Picpus, 75012, Paris, France

Corresponding Author: vivien.deshaies@aphp.fr

Avec l'essor de la médecine de précision et l'utilisation en routine des méthodes de séquençage haut-débit, le volume de données à interpréter croit rapidement. L'utilisation de solutions logicielles dédiés à cette phase d'interprétation devient indispensable afin de garantir un délai de rendu au patient.

Au sein des 39 hôpitaux de l'AP-HP, les solutions logicielles utilisées sont très diverses. La plupart des hôpitaux utilisent notamment les service d'entreprise privées sur un modèle SaaS (Software as a Service), parfois peu respectueuses des règles de gestion des données de patients en France. D'autres hôpitaux utilisent des solutions logicielles installées sur poste de travail, ne permettant pas une gestion centralisée des variants pour l'ensemble des équipes de diagnostic moléculaire de l'AP-HP et ne favorisant ainsi pas le partage de l'expertise au sein de l'institution.

Afin d'adresser ces différentes problématiques, nous avons développé Leaves, une solution logicielle déployée et maintenue centralement au niveau de la plateforme bioinformatique MOABI et disponible pour l'ensemble des équipes de l'AP-HP. Son rôle est d'une part de proposer une solution efficace pour l'interprétation des résultats et d'autre part de favoriser le partage d'expertise entre professionnels tout en homogénéisant les méthodes d'interprétation et ainsi d'améliorer le service rendu aux patients.

Leaves est une interface web développée en python et javascript avec entre autre les frameworks Flask, sqlalchemy, D3.js, jQuery et Vue.js. Cette application permet une utilisation multi-utilisateurs et multi-équipes. Chaque équipe peux réaliser ses interprétations pour l'intégralité de ses panels en fonction des pathologies étudiées, qu'elles soient du domaine du cancer ou du domaine des maladies rares.

Leaves rassemble une grande diversité de banques publiques d'annotations issues pour partie de dbNSFP [1][2] et snpeff [3]. Ces informations sont utilisables par l'utilisateur pour filtrer et trier ses résultats. Les filtres peuvent soit être appliqués dynamiquement par l'utilisateur directement au niveau du tableau de résultats, soit être intégrés sous forme de pipelines de filtres qui seront ensuite lancés automatiquement à chaque insertion de nouvelle données et donneront ainsi un accès direct au données pré-filtrées.

Enfin, Leaves permet le partage d'interprétations et de commentaires entre utilisateurs. Les interprétations effectuées dans Leaves suivent les critères de classification recommandés par l'American College of Medical Genetics and Genomics (ACMG) [4] et l'Association Nationale des Praticiens de Génétique Moléculaire (ANPGM). L'utilisation de ces critères permet de tendre vers une normalisation des interprétations entre praticiens et une structuration forte de la donnée permettant une réutilisation ultérieure.

Bibliographie

1: Liu X, Jian X, and Boerwinkle E., dbNSFP: a lightweight database of human non-synonymous SNPs and their functional predictions., 2011
2: Liu X, Jian X, and Boerwinkle E., dbNSFP v2.0: A Database of Human Non-synonymous SNVs and Their Functional Predictions and Annotations., 2013
3: Cingolani P, Platts A, Wang le L, Coon M, Nguyen T, Wang L, Land SJ, Lu X, Ruden DM., A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of Drosophila melanogaster strain w1118; iso-2; iso-3., 2012
4: Richards Sue, Aziz Nazneen, Bale Sherri, Bick David, Das Soma, Gastier-Foster Julie, Grody Wayne W., Hegde Madhuri, Lyon Elaine, Spector Elaine, Voelkerding Karl, Rehm Heidi L., Standards and guidelines for the interpretation of sequence variants: a joint consensus recommendation of the American College of Medical Genetics and Genomics and the Association for Molecular Pathology, 2015

Linking structural and evolutionary information using MIToS.jl

Diego Javier Zea[1]
LCQB UMR 7238 CNRS, IBPS, Sorbonne Université, 7 Quai Saint-Bernard, 75005, Paris, France

Corresponding author: diegozea@gmail.com

1 Introduction

MIToS is a *Julia* package for analyzing protein sequence and structure, with the main focus on coevolutionary analysis [1]. However, its utilities go beyond the calculation of covariation scores in multiple sequence alignments. *MIToS* is a flexible suite that has been used to measure residue conservation, to deal with protein structures in homology modelling and molecular dynamics pipelines, to perform structural alignment of tertiary and quaternary structures, etc. *MIToS* allows to access the power of *Julia*, a high-level programming language for scientific computing with a close to *C* performance [2].

MIToS defines functions and types for dealing with multiple sequence alignments, parsing protein structures, determine inter-residue contacts and interactions, mapping information between sequence and structure using *SIFTS* [3] and many other tasks. Their modules allow to write and run an entire protein sequence and structure analysis pipeline in a single programming language. Julia performance and easy to use parallelism allow us to run these analyses on large datasets and to test multiple hypotheses, parameter combinations, etc. As a result, it was used to create new knowledge about the relation between the evolutionary signals and the change of protein structures through evolution [4].

The software is totally implemented in Julia and supported for Linux, OS X and Windows. It's freely available on GitHub under MIT license: *https://github.com/diegozea/MIToS.jl*

2 Demonstration

From the multiple possible applications that *MIToS* allows, the demonstration is going to focus on the mapping between sequence and structure. This is a very common task for linking structural information coming from *PDB* and evolutionary information calculated from multiple sequence alignments. *MIToS* makes this task easier by keeping the mapping information in the multiple sequence alignment annotations. In this way, it is possible to track residue positions, even after deleting alignment columns. Also, the ability of *MIToS* to parse *SIFTS* files allows to access their residue level mapping between PDB and other databases, e.g. *UniProt*. Both things together allow the correct mapping of sequence and structure without performing error-prone pairwise alignments.

The demonstration will be done with *Jupyter* notebooks. Those notebooks are going to be available at *GitHub* with a *Binder* set up to allow attendants to execute the code.

Acknowledgements

That work was supported by Sorbonne Université, CONICET, FIL and PICT 2014-1787.

References

[1] Diego J Zea, Diego Anfossi, Morten Nielsen, and Cristina Marino-Buslje. Mitos. jl: mutual information tools for protein sequence analysis in the julia language. *Bioinformatics*, 33(4):564–565, 2016.

[2] Jeff Bezanson, Alan Edelman, Stefan Karpinski, and Viral B Shah. Julia: A fresh approach to numerical computing. *SIAM review*, 59(1):65–98, 2017.

[3] Jose M Dana, Aleksandras Gutmanas, Nidhi Tyagi, Guoying Qi, Claire O'Donovan, Maria Martin, and Sameer Velankar. Sifts: updated structure integration with function, taxonomy and sequences resource allows 40-fold increase in coverage of structure-based annotations for proteins. *Nucleic acids research*, 47(D1):D482–D489, 2018.

[4] Diego Javier Zea, Alexander Miguel Monzon, Gustavo Parisi, and Cristina Marino-Buslje. How is structural divergence related to evolutionary information? *Molecular phylogenetics and evolution*, 127:859–866, 2018.

Omics Visualizer: a Cytoscape App to visualize omics data

Marc Legeay[1], Nadezhda T. Doncheva[1,2], John H. Morris[3] and Lars J. Jensen[1]
[1] Novo Nordisk Foundation Center for Protein Research, University of Copenhagen, Denmark
[2] Center for Non-Coding RNA in Technology and Health, Department of Veterinary and Animal Sciences, University of Copenhagen, Denmark
[3] Resource for Biocomputing, Visualization and Informatics, University of California, San Francisco, CA 94143, USA

Corresponding author: lars.juhl.jensen@cpr.ku.dk

1 Abstract

Omics Visualizer is a Cytoscape app that allows users to import data tables with multiple rows referring to the same network node and to visualize such data onto networks. This is particularly useful for visualizing post-translational modification sites or peptides identified in proteomics studies as well as data measured under multiple conditions. The app is freely available at: http://apps.cytoscape.org/apps/omicsvisualizer.

2 Introduction

Cytoscape [1] is an open-source software used to analyze and visualize networks. In addition to being able to import networks from a variety of sources, Cytoscape allows users to import tabular node data and visualize it onto networks. Unfortunately, such data tables can only contain one row of data per node, whereas omics data often have multiple rows for the same gene or protein, representing different post-translational modification sites, peptides, splice isoforms, or conditions. However, Cytoscape has an API that allows developers to make apps that extend its functionality. Here, we present a new app, Omics Visualizer, that allows users to import data tables with several rows referring to the same node and visualize such data.

3 Description of the App

Omics Visualizer enables users to import a data table, connect it to one or more networks, and visualize the connected data onto networks. If the user does not provide a network, Omics Visualizer retrieves a network from the STRING database [2] using the Cytoscape stringApp [3].

The Omics Visualizer table import mimics the Cytoscape default import process: it handles text and spreadsheet files, the user can select the columns to import and modify the auto-detected type for each. To connect a table with a network, the user must select the key columns from the node table and from the data table. Omics Visualizer gives the possibility to represent numerical data from the table onto the network in two ways: either with a pie chart in the node, or with a donut chart around the node. A slice of the chart is a row from the table, and the color represents the numerical value. The charts are drawn by the enhancedGraphics app [4].

Acknowledgments

This works was funded by the Novo Nordisk Foundation (NNF14CC0001).

References

[1] Paul Shannon, Andrew Markiel, Owen Ozier, Nitin S. Baliga, Jonathan T. Wang, Daniel Ramage, Nada Amin, Benno Schwikowski, and Trey Ideker. Cytoscape: A Software Environment for Integrated Models of Biomolecular Interaction Networks. *Genome Research*, 13(11):2498–2504, 2003.

[2] Damian Szklarczyk, Annika L. Gable, David Lyon, Alexander Junge, Stefan Wyder, Jaime Huerta-Cepas, Milan Simonovic, Nadezhda T. Doncheva, John H. Morris, Peer Bork, Lars J. Jensen, and Christian von Mering. STRING v11: protein-protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets. *Nucleic Acids Research*, 47(D1):D607–D613, 2019.

[3] Nadezhda T. Doncheva, John H. Morris, Jan Gorodkin, and Lars J. Jensen. Cytoscape StringApp: Network Analysis and Visualization of Proteomics Data. *Journal of Proteome Research*, 18(2):623–632, 2019.

[4] John H. Morris, Allan Kuchinsky, Thomas E. Ferrin, and Alexander R. Pico. enhancedGraphics: a Cytoscape app for enhanced node graphics. *F1000Research*, 3:147, 2014.

Reconstruction of Transcript phylogenies using PhyloSofS

Adel Ait-hamlat[1], Lélia Polit[1], Antoine Labeeuw[1], Diego Javier Zea[1], Hugues Richard[1] and Elodie Laine[1]
Laboratory of Computational and Quantitative Biology, 4 Place Jussieu, 75005, Paris, France

Corresponding author: elodie.laine@upmc.fr

1 PhyloSofS

Alternative splicing (AS) has the potential to greatly expand the proteome in eukaryotes, by producing several transcript isoforms from the same gene. AS has been associated with multiple biological functions [1,2], and its deregulation has been associated with the development of various diseases [3].

We developed PhyloSofS, a fully automated computational tool that infers plausible evolutionary scenarios explaining a set of transcripts observed in several species and models the three-dimensional structures of the produced isoforms [4]. The method provides a mean to address unresolved questions linked to alternative splicing events. PhyloSofS can help us identify alternative splicing events inducing substantial conformational rearrangements or even fold changes and discovering new therapeutic targets (isoforms) and can also shed light on the evolutionary paths leading to functional innovation.

In this demonstration we are going to show the pipeline of PhyloSofS, and what results it can generate. The algorithm takes a binary gene tree for a set of species and their ensemble of transcripts as an input. A forest of phylogenetic trees describing plausible evolutionary scenarios that can explain the observed transcripts is reconstructed using the maximum parsimony principle. Then, PhyloSofS' phylogenetic reconstruction algorithm provides the user with the evolutionary history of the isoforms.

On the second part of the demo, we will focus on the structural aspects of PhyloSofS. The pipeline, based on HH-suite, creates a structure by homology modelling of every isoforms in the gene family, then annotates it. This modelling of every isoforms, coupled with the transcripts annotations, could help us to get insight into the molecular mechanisms underlying AS-induced functional changes.

PhyloSofS has been used on 12 gene families and will be used on the whole human proteome in the future.

PhyloSofS is open-source and is freely available at https://github.com/PhyloSofS-Team/PhyloSofS.

References

[1] Olga Kelemen, Paolo Convertini, Zhaiyi Zhang, Yuan Wen, Manli Shen, Marina Falaleeva, and Stefan Stamm. Function of alternative splicing. *Gene*, 514(1):1 – 30, 2013.

[2] Stephen J Bush, Lu Chen, Jaime M Tovar-Corona, and Araxi O Urrutia. Alternative splicing and the evolution of phenotypic novelty. *Philosophical transactions of the Royal Society of London. Series B, Biological sciences*, 372(1713):20150474, 02 2017.

[3] Amanda J Ward and Thomas A Cooper. The pathobiology of splicing. *The Journal of Pathology*, 220(2):152–163.

[4] Adel Ait-hamlat, Lélia Polit, Hugues Richard, and Elodie Laine. Transcripts evolutionary conservation and structural dynamics give insights into the role of alternative splicing for the jnk family. *bioRxiv*, 2017.

S3A: A Scalable and Accurate Annotated Assembly Tool for Gene Assembly

Laurent DAVID[1], Riccardo VICEDOMINI[1,2] , Hugues RICHARD[1] and Alessandra CARBONE[1,3]

[1] Sorbonne Université, CNRS, IBPS, Laboratoire de Biologie Computationnelle et Quantitative (LCQB), UMR 7238, Paris 75005, France
[2] Sorbonne Université, Institut des Sciences, du Calcul et des Données, Paris 75005, France
3 Institut Universitaire de France, Paris 75005, France

Corresponding Author: alessandra.carbone@lip6.fr, hugues.richard@upmc.fr

1 Introduction

Next-generation sequencing of environmental samples aims at studying microbial communities. It is commonly followed by a functional annotation of the predicted coding regions in order to describe the community's metabolic activities. In metagenomics, annotation is hampered for shorter sequences, thus making sequence assembly a prerequisite for any improvement. In this context, a good-quality assembler is necessary, as it increases the actual length of coding regions. The sheer size of the metagenomic datasets requires huge time and memory ressources when doing de novo metagenome assembly. Thus, several strategies have been proposed to perform a targeted assembly, based on preliminary protein domain annotation followed by domain guided assembly. S3A is a domain based targeted assembler specifically designed to maintain good accuracy while controlling running time complexity.

2 S3A algorithm and key features

S3A exploits reads annotation as a first indicator of read overlaps, and then applies string-based filtering to clustered annotated reads in order to find bona fide overlaps. It constructs an Overlap Layout Consensus (OLC) Graph, a directed graph where each node corresponds to a read and each edge to an overlap between two reads. The goal behind the construction of the OLC graph is to perform a depth-first graph traversal leading to the reconstruction of contigs, that is consensus regions of DNA assembled from the sets of overlapping reads.

The design of the S3A algorithm is driven by the motivation of reducing time complexity while retaining the highest accuracy possible in assembly reconstruction. The first main idea is to avoid useless comparisons by separating reads annotated with different domains. Ordering reads by their matching position on the domain further improves the general algorithm performance and greatly lowers the number of comparisons. The second idea is, for each pair of domain-overlapping reads, to compute two fast and complementary metrics : the *longest matching common substring length* (*lms*) and the *identity percentage* (*ip*). These metrics give a strong overlapping confidence measure that is both complementary and much faster than computing an edit distance. Another advantage of using these metrics is that they allow for a tailored graph trimming which is independent on the sequencing technology used and helps reducing graph complexity. Moreover, *lms* is used to select the most reliable *transitive edges* (which are edges connecting nodes that have an alternative path joining them, and help guiding the traversal by giving a stronger evidence that two given nodes belong to the same coding region), and to resolve ambiguous cases in the absence of *transitive edges*.

3 Software

The S3A assembler takes as input a set of reads in fasta format, that it automatically annotates with Open Reading Frames (e.g. by FragGeneScan [1]) and domains (e.g. by MetaCLADE [2] or HMMER [3]). Its output is a text file containing the list of reads assembled (contigs). S3A default parameter values are optimized to obtain best results, but can be modified by the user by command line options. In summary S3A enables the rapid profiling of a predefined set of domain on a metagenomic sample, while maintaining good accuracy and a reasonable running time. S3A is available at http://www.lcqb.upmc.fr/S3A_ASSEMBLER/ .

References

[1] Rho M, Tang H, Ye Y. FragGeneScan: predicting genes in short and error-prone reads. *Nucleic Acids Res* 38: e191 (2010).
[2] Ugarte A.,Vicedomini R., Bernardes J., Carbone A. A multi-source domain annotation pipeline for quantitative metagenomic and metatranscriptomic functional profiling. *Microbiome*, 6:149, 2018.
[3] Finn, R.D., Clements, J. & Eddy, S.R. HMMER web server: interactive sequence similarity searching. *Nucleic Acids Res*. 39, W29-W37 (2011).

T1TAdb: the database of Type I Toxin-Antitoxin systems

Nicolas J. TOURASSE[1] and Fabien DARFEUILLE[1]

[1] ARNA laboratory, INSERM U1212, CNRS UMR 5320, University of Bordeaux, 146 rue Léo Saignat, 33076, Bordeaux, France

Corresponding Author: fabien.darfeuille@inserm.fr

Toxin-antitoxin (TA) systems are small genetic loci found in most bacterial genomes including those of pathogens. They are usually composed of two adjacent genes: a stable toxin and a labile antitoxin, whose depletion rapidly leads to death or growth arrest (see [1] for a recent review). Six types of TA systems have been described so far depending on the nature and mode of action of the antitoxin. While the toxin is always a protein, the antitoxin can be either a protein (types II, IV, V and VI) or an RNA (types I and III). A type I TA system consists of an mRNA coding for a small peptide (20-60 amino acids) that is toxic to the host cell and an antisense noncoding small RNA (asRNA; 60-200 nucleotides) that serves as a counteracting antitoxin to prevent the synthesis of its cognate toxin by directly basepairing to the mRNA.

A few type I TA systems have been experimentally characterized, and hundreds have been identified by bioinformatic analyses [2,3]. Most of them are not annotated in genome records. Therefore, there is an urgent need for a central repository. In this work, we have thus built a database for type I TA systems, T1TAdb, that gathers all described and predicted loci. In addition, as the majority of loci have been predicted solely on the basis of the toxin peptide sequence, we devised a procedure to annotate the mRNA and asRNA coordinates. Both genes were identified based on key determinants of the secondary structure that are important for their expression and/or activation, such as sequestration of the ribosome-binding site, terminator stems, and long-distance 5'-3' interactions, as well as the organization of the two genes within the locus. More specifically, the genomic regions defining the mRNAs and asRNAs were predicted using RNAMotif [4] and RNASurface [5], respectively. RNAMotif identifies regions that can adopt a predefined secondary structure, while RNASurface predicts regions that are structurally more stable than the rest of the genome.

T1TAdb is implemented as a relational database in PostgreSQL and the graphical web interface was developed using the PERL Catalyst framework, along with the PERL Template Toolkit templating system. The database is manually curated and provides tools for viewing, searching, and comparing sequence, structure, and genomic data on type I TA systems, and thus may be a valuable resource to gain a better understanding of their distribution, evolution, and function. T1TAdb currently contains ~2,000 loci from ~500 genomes described in previous studies [2,3] and is freely available at https://d-lab.arna.cnrs.fr/t1tadb.

Acknowledgements

The web server hosting T1TAdb is provided by the ODS Web Hosting service of CNRS. We thank Dr. Stéphane Thore and Dr. Sébastien Fribourg for providing additional computing power.

References

1. Alexander Harms, Ditlev E. Brodersen, Namiko Mitarai, and Kenn Gerdes. Toxins, Targets, and Triggers: An Overview of Toxin-Antitoxin Biology. *Molecular Cell*, (70):768-784, 2018.
2. Hélène Arnion, Dursun N. Korkut, Sara Masachis Gelo, Sandrine Chabas, Jérémy Reignier, Isabelle Iost, and Fabien Darfeuille. Mechanistic insights into type I toxin antitoxin systems in *Helicobacter pylori:* the importance of mRNA folding in controlling toxin expression. *Nucleic Acids Research*, (45):4782-4795, 2017.
3. Elizabeth M. Fozo, Kira S. Makarova, Svetlana A. Shabalina, Natalya Yutin, Eugene V. Koonin, and Gisela Storz. Abundance of type I toxin-antitoxin systems in bacteria: searches for new candidates and discovery of novel families. *Nucleic Acids Research*, (38): 3743–3759, 2010.
4. Thomas J. Macke, David J. Ecker, Robin R. Gutell, Daniel Gautheret, David A. Case, and Rangarajan Sampath. RNAMotif, an RNA secondary structure definition and search algorithm. *Nucleic Acids Research*, (29): 4724–4735, 2001.
5. Ruslan A. Soldatov, Svetlana V. Vinogradova, and Andrey A. Mironov. RNASurface: fast and accurate detection of locally optimal potentially structured RNA segments. *Bioinformatics*, (30): 457–463, 2014.

Platform session

AskOmics: a user-friendly interface to Semantic Web technologies for integrating local datasets with reference resources

Xavier Garnier[1], Anthony Bretaudeau[1,2], Fabrice Legeai[1,2], Anne Siegel[1]and Olivier Dameron[1]
[1] Univ Rennes, CNRS, Inria, IRISA - UMR 6074, F-35000 Rennes, France
[2] Institut de Génétique, Environnement et Protection des Plantes (IGEPP) - Institut national de la recherche agronomique (INRA) : UMR1349, Agrocampus Ouest - Agrocampus Ouest, UMR1349 IGEPP, F-35 042 Rennes, France

Corresponding author: xavier.garnier@irisa.fr

1 Introduction

The study of biological mechanisms require the production of large and heterogeneous datasets. These *omics* datasets are obtained routinely into labs, and are also available from public databases. Each of them have their own format and linking them require lot of time to link them. To ease integration of this data, we have developed AskOmics, a web tool designed to integrate heterogeneous biological data, and query them using a user-friendly interface. The software uses the semantic web technologies in order to homogenize data. AskOmics is used by several research team to analyze genomics, transcriptomics and pathways data. Software development is still ongoing and new features are added as new versions are released. Next version (AskOmics 3) will brings a new set of features.

2 AskOmics, integration and query using the semantic web technologies

AskOmics is a web software that uses the semantic web technologies (RDF/SPARQL) to integrate multiple data formats, and query them through a user-friendly interface. During data integration, user provides input files in common formats (CSV, GFF and BED). AskOmics internally generates the corresponding RDF triples and load them into a triplestore. Two kinds of information are generated, the *content*, corresponding to the raw data, and the *abstraction*, which describe how the raw data are organized and interlinked.

The query interface is composed of a dynamic graph that uses the generated abstraction to represent the entities integrated. Users interact with the graph to build a complex query linking several datasets integrated in Askomics. When the query is built, AskOmics internally converts the graph into a SPARQL query and use it to interrogate the triplestore. Results are returned to the web interface and can be downloaded by the user.

AskOmics source code is available under AGPL3 licence at https://github.com/askomics/askomics. The GenOuest bioinformatics platform hosts a sandbox instance at https://askomics.genouest.org.

3 Ongoing work

We are currently developing version 3 of AskOmics (https://github.com/xgaia/flaskomics). This version will bring new features: the possibility to generate an AskOmics abstraction from RDF data, the hierarchy management between entities, and the implementation of federated queries against external endpoints, such as uniprot. Users will be able to link their data to large existing databases without having to import them locally.

This version has a new graphical user interface, build with React, a Javascript library made for building user interface. this change will provide a more modern and maintainable interface. The Python API is being refactored using Flask framework and Celery task queue for better performance by reducing the number of calls and making asynchronous call for long tasks such as data integration.

This new version is splitted into several micro services, *AskOmics*, a Python API and a Javascript interface. *Celery*, a task queue, to execute long task asynchronously, *Redis*, a worker, dependency of Celery, *Virtuoso*, a triplestore, for storing RDF and accepting SPARQL queries, and *Nginx*, a web proxy for url redirecting to the services. All these microservices are provided as docker containers. AskOmics deployment and upgrading is therefore very easy.

DevOps bioinformatics services with Docker, GitLab CI, and Kubernetes

Bryan Brancotte[1], Thomas Menard[2] and Hervé Menager[1]

[1] Bioinformatics and Biostatistics Hub of the C3BI, Institut Pasteur, Paris, France
[2] Direction des système de l'information, Institut Pasteur, Paris, France

Corresponding Author: bryan.brancotte@pasteur.fr

Summary

When proposing software, reliability is critical for adoption. Tests prevent code regression and reassure users that it will behave as expected, documentation allows to actually use it. For a web service, being highly available [1] and reactive is mandatory.

Tests are not always easily reproducible, it become even harder when multiple software dependencies are involved. The deployment and monitoring of an application can be documented, but is rarely automated and therefore most often time consuming.

DevOps [2] is a set of practices which help tackle these issues. Unit tests are reproducible and prevent regression. Using docker containers, the tests and installation are automated, documented and reproducible. Kubernetes proposes multiple environments, highly available services with monitoring and failover, scalability and load balancing. On top of them, GitLab offers a Continuous Integration which verifies at each commit that tests are passed, and automated deployment.

In this poster we present how this "DevOps" infrastructure is used at the Institut Pasteur to develop and deploy new bioinformatics services, with an overview of the architecture and its usage, as well as concrete examples.

Keywords

Docker
Kubernetes
GitLab
Continuous integration
Continuous delivery
DevOps
Web services
Reproducible science
Quality software

References

[1] Schultheiss, Sebastian & Münch, Marc-Christian & D Andreeva, Gergana & Rätsch, Gunnar. (2011). Persistence and Availability of Web Services in Computational Biology. PloS one. 6. e24914. 10.1371/journal.pone.0024914.
[2] L. Zhu, L. Bass and G. Champlin-Scharff, "DevOps and Its Practices," in IEEE Software, vol. 33, no. 3, pp. 32-34, May-June 2016. doi: 10.1109/MS.2016.81

IFB-Biosphère, Portail pour le Déploiement de Services Bioinformatiques sur une Fédération de Clouds

Jonathan LORENZO[1], Matéo BOUDET[2], Jean-François GUILLAUME[3], Efflam LEMAIILET[2], Stéphane DELMOTTE[4], Olivier SALLOU[2], Bruno SPATARO[4], Jérôme PANSANEL[5], Hervé GILQUIN[6], Olivier COLLIN[2],Christophe BLANCHET[1]

[1] *CNRS, UMS 3601 ; Institut Français de Bioinformatique, IFB-core, F-91000 Evry, France*

[2] *GenOuest, Univ Rennes, Inria, CNRS, IRISA F-35000 Rennes*

[3] *BiRD, UMR_S1087/UMR_C6291, LS2N UMR 6004, Université de Nantes, F-44007 NANTES, France*

[4] *PRABI-LBBE*

[5] *Université de Strasbourg, CNRS UMR7178 IPHC, F-67037 Strasbourg, France*

[6] *CNRS, UMR 5669, PSMN (Pôle Scientifique de Modélisation Numérique) ENS de Lyon, Lyon*

Corresponding Author: christophe.blanchet@france-bioinformatique.fr

L'Institut Français de Bioinformatique (IFB) propose différents services pour le traitement des données des sciences de la vie. Une partie de cette offre de services est basée sur un cloud académique, mettant à disposition de la communauté les très nombreux logiciels et collections de données biologiques permettant d'analyser les données expérimentales produites couramment. L'infrastructure de cloud bioinformatique de l'IFB est distribuée entre des plates-formes régionales et le nœud national, sous la forme d'une fédération de clouds, IFB-Biosphère.

Le portail Biosphère[1] fournit plusieurs interfaces pour simplifier l'usage de l'infrastructure cloud distribuée de l'IFB :

- le catalogue RAINBio des « appliances cloud », qui référence les environnements basés sur des machines virtuelles (VM) prêtes à être déployées en un clic, dimensionnées pour différentes tâches bioinformatiques,
- un tableau de bord qui permet à chaque usager de gérer ses déploiements dans le cloud IFB-Biosphère, qu'ils reposent sur une seule ou plusieurs machines virtuelles,
- un centre de données qui recense les banques de données publiques, disponibles dans les clouds IFB-Biosphère. Ces banques de données, accessibles en mode fichier, sont montées directement dans les machines virtuelles des utilisateurs.

Les appliances bioinformatiques du cloud IFB-Biosphère sont disponibles en différents formats pour différentes thématiques, permettant aux scientifiques, biologistes et bioinformaticiens, de choisir le plus approprié pour leurs analyses.

Dans notre démonstration, nous déploierons diverses appliances proposant une interface scientifique de haut-niveau reposant sur des portails web comme Rstudio et Jupyter Notebook, et/ou des interfaces graphiques (GUI) à travers un bureau virtuel à distance comme ImageJ et Cytoscape. Ces environnements virtuels de recherche sont référencés dans le catalogue RAINBio et déployables en un clic avec la configuration type définie par leurs développeurs. Ils peuvent aussi être adaptés par l'utilisateur suivant ses besoins sans interférer avec les autres usagers avec des outils technologiques comme `conda`, `docker` ou `ansible` pour le déploiement automatisé de logiciels. Ensuite, pour chaque déploiement, un lien permettant d'accéder à l'interface est disponible dans le tableau de bord , et permet de gérer les déploiements en cours (vue détaillée et suppression).

[1] https://biosphere.france-bioinformatique.fr

IFB-Biosphère, Services Cloud d'Analyse des Données des Sciences de la Vie

Christophe BLANCHET[1], Olivier COLLIN[2], Matéo BOUDET[2], Stéphane DELMOTTE[3], Hervé GILQUIN[4], Jean-François GUILLAUME[5], Efflam LEMAIILET[2], Jonathan LORENZO[1], Jérôme PANSANEL[6], Olivier SALLOU[2], Bruno SPATARO[3]

[1] *CNRS, UMS 3601 ; Institut Français de Bioinformatique, IFB-core, F-91000 Evry, France*

[2] *GenOuest, Univ Rennes, Inria, CNRS, IRISA F-35000 Rennes*

[3] *PRABI-LBBE*

[4] *CNRS, UMR 5669, PSMN (Pôle Scientifique de Modélisation Numérique) ENS de Lyon, Lyon*

[5] *BiRD, UMR_S1087/UMR_C6291, LS2N UMR 6004, Université de Nantes, F-44007 NANTES, France*

[6] *Université de Strasbourg, CNRS UMR7178 IPHC, F-67037 Strasbourg, France*

Corresponding Author: christophe.blanchet@france-bioinformatique.fr

L'Institut Français de Bioinformatique (IFB) propose différents services pour le traitement des données des sciences de la vie. Une partie de cette offre de services est basée sur un cloud académique, mettant à disposition de la communauté les très nombreux logiciels et collections de données biologiques permettant d'analyser les données expérimentales produites couramment. L'infrastructure de cloud bioinformatique de l'IFB est distribuée entre des plates-formes régionales et le nœud national, sous la forme d'une fédération de clouds, IFB-Biosphère. Le portail Biosphère[1] fournit plusieurs interfaces pour simplifier l'usage de l'infrastructure : le catalogue RAINBio des appliances bioinformatiques, un tableau de bord des VMs et un pour les données.

La fédération de clouds IFB-Biosphère a été initiée en 2016, et comporte actuellement plus de 5 200 cœurs de calcul et 26 téraoctets (To) de mémoire. Ces ressources sont réparties entre 5 sites : GenOuest, PRABI-LBBE, BiRD, BIstrO et le nœud national IFB-core. Certains de ces clouds fonctionnent depuis le début des années 2010, et 5 autres plates-formes de l'IFB souhaitent raccorder leur cloud à la fédération. L'infrastructure cloud IFB-Biosphère est accessible à l'ensemble de la communauté des sciences de la vie, avec un quota de ressources de base, extensible selon différents critères. Ces ressources peuvent aller de 1 vCPU-2 Go RAM à 128 vCPU-3 To RAM pour une seule machine virtuelle, jusqu'à des centaines ou milliers de cœurs avec des centaines de Go ou plusieurs To de mémoire dans de nombreuses machines virtuelles.

Les appliances bioinformatiques (environnements basés sur des machines virtuelles) sont disponibles en différents formats pour différentes thématiques. Il y a actuellement 30 environnements modèles, développés par les membres de l'IFB, référencés dans le catalogue RAINBio. Ces appliances proposent de nombreux outils courants en bioinformatique, modules R… très utilisés pour l'analyse de données par exemple en génomique, bio-imagerie, réseaux métaboliques, écologie microbienne, protéomique ou métabolomique. Certains environnements fournissent des outils technologiques comme `conda` (avec les canaux bioconda et R pré-configurés), `docker` pour les conteneurs, ou `ansible` pour le déploiement automatisé de logiciels. D'autres environnements proposent des interfaces web (comme Rstudio, Jupyter Notebook ou Galaxy), ou des interfaces graphiques (GUI) à travers un bureau virtuel à distance. Ces environnements virtuels de recherche se déploient avec la configuration type définie par leurs développeurs, mais tous peuvent être adaptés par l'utilisateur suivant ses besoins sans interférer avec les autres usagers.

Le cloud IFB-Biosphère est ainsi utilisé pour des analyses scientifiques intensives (jusqu'à 4 000 cœurs de calcul) et par de nombreuses sessions de formation, écoles scientifiques, cursus de masters universitaires, workshops ou hackathons, dont certains depuis plusieurs années.

Remerciements

IFB est soutenu par l'ANR (ANR-11-INBS-0013) et par les organismes CNRS, INRA, Inserm, CEA et Inria.

[1] https://biosphere.france-bioinformatique.fr

A Shiny and Galaxy interactive software for multi-source data analysis

Etienne CAMENEN[1,2], Arnaud GLOAGUEN[3], François-Xavier LEJEUNE[1], Ivan MOSZER[1] and Arthur TENENHAUS[1,3]

[1] Institut du Cerveau et de la Moelle épinière, Inserm U 1127, CNRS UMR 7225, Sorbonne Université, F-75013, Paris, France

[2] Institut Français de Bioinformatique, CNRS UMS 3601, F-91057, Evry, France

[3] Laboratoire des Signaux et Systèmes (L2S), CentraleSupélec, Université Paris-Saclay, F-91190, Gif-sur-Yvette, France

Corresponding Author: etienne.camenen@icm-institute.org

Regularized Generalized Canonical Correlation Analysis (RGCCA) is a statistical framework for multiblock data analysis and encompasses as special cases a remarkably large number of multiblock components methods [1,2]. From an application viewpoint, this method is currently limited to “expert” users through the RGCCA R package [3]. We propose to develop interactive and ergonomic interfaces for biologists to facilitate the parameterization and visualize the outputs of RGCCA analyses “on the fly”, through the Galaxy and Shiny environments. The usefulness and versatility of these interfaces are evaluated on multi-source biological data sets to identify biomarkers of Parkinson’s disease severity.

As this software is designed for non-statisticians, all the tuning parameters of RGCCA are predefined to default values. The sole required step is related to the construction of the multiblock data set where the variables that compose each block have to be defined before the analysis. Once the blocks are specified, the RGCCA analysis is automatically launched and the visualization of the results is available through several graphical outputs. An “advanced mode” allows users more familiarized with RGCCA to tune the parameters and to navigate through specific outputs of the RGCCA analysis. The interactive graphical representations help clinicians identifying and visualizing subsets of variables within each block that may explain the links between blocks. These sets of candidate variables can eventually be associated with a clinical response or groups of subjects.

Work in progress includes: (i) adding new statistical features such as the automatic estimation of the tuning parameters of RGCCA (based on cross-validation or permutation), (ii) handling block-wise missing values, (iii) adding functions for multigroup [5] and multiway analysis [6], (iv) integrating the Shiny interface in the next release of the RGCCA package and (v) integrating the Shiny developments into a Galaxy wrapper to combine the benefits of both types of environments and facilitate the integration of our tools into the French Galaxy community Workflow4Metabolomics (W4M).

Acknowledgements

EC is funded by the Institut Français de Bioinformatique (ANR-11-INSB-0013) in the framework of the pilot project “IntegrParkinson”. This work was also partly supported by the IHU-A-ICM program ANR-10-IAIHU-06.

References

[1] Tenenhaus A and Tenenhaus M. Regularized generalized canonical correlation analysis. *Psychometrika*, 76:257-284, 2011.

[2] Tenenhaus M, Tenenhaus A and Groenen PJF. Regularized generalized canonical correlation analysis: A framework for sequential multiblock component methods. *Psychometrika*, 82(3):737–777, 2017.

[3] Tenenhaus A and Guillemot V. RGCCA Package, 2017. http://cran.project.org/web/packages/RGCCA/index.html

[4] Garali I, Adanyeguh I, Ichou F, Perlbarg V, Seyer A, Colsch B, Moszer I, Guillemot V, Durr A, Mochel F, Tenenhaus A. A strategy for multimodal data integration: application to biomarkers identification in spinocerebellar ataxia. *Briefings in Bioinformatics*, 19(6): 1356-1369, 2018.

[5] Tenenhaus A and Tenenhaus M, Regularized generalized canonical correlation analysis for multiblock or multigroup data analysis. *European Journal of Operational Research*, 238(2), 391-403, 2014.

[6] Tenenhaus A, Le Brusquet L and Lechuga G. Multiway Regularized Generalized Canonical Correlation Analysis. 47ème Journées de Statistique, Lille, France, 2015.

Vers le déploiement continu d'infrastructures de calculs pour la bioinformatique

David BENABEN[1,2], Nicole Charrière[3], Gildas LE CORGUILLÉ[4], Julien SEILER[5] and Guillaume SEITH[5]

[1] CBiB, Université de Bordeaux, 142 rue Léo Saignat, 33076 Bordeaux, France, [2] INRA, UMR 1332, Biologie du Fruit et Pathologie, CS20032 Villenave d'Ornon, France, [3] IFB/Institut Français de Bioinformatique, CNRS UMS 3601, IFB-Core, Génoscope, 91057, ÉVRY, France, [4] CNRS/Sorbonne Université, FR2424, ABiMS, Station Biologique, 29680, Roscoff, France, [5] IGBMC, 1 rue Laurent Fries, 67404, Illkirch, France

Corresponding Authors: lecorguille@sb-roscoff.fr, julien.seiler@igbmc.fr

Dans le cadre du plan d'action 2018-2021 et de la mise en place d'un environnement réparti pour le traitement des données (NNCR : National Network of Computational Resources), l'IFB a déployé, en complément de l'infrastructure Cloud, une ressource de calcul centrale type HPC (High Performance Computer): l'IFB Core Cluster. Cette ressource est hébergée à l'IDRIS et offre une capacité de 2000 cœurs et 1 Po de stockage. Le Core Cluster intègre des composants pour le calcul (SLURM…), le stockage (NFS, stockage MooseFS), pour la virtualisation (ProxMox, VMWare) et met à disposition des environnements logiciel (via Conda, Singularity ou des portails Web). Il est élaboré par un collectif d'une dizaine d'ingénieurs de plateformes régionales de l'ensemble du réseau de plateformes IFB (« *mutualised task force* ») qui dédient un pourcentage de leur temps à l'élaboration de ce projet commun. Toutes les procédures d'installation reposent sur des procédures d'Intégration Continue (CI) établies en commun (recettes Ansible, packages Conda …) qui permettent de reproduire les mêmes environnements logiciels sur d'autres infrastructures comme les clusters régionaux IFB. Cette organisation permet aussi à tout un chacun de participer à l'administration du cluster sans droit root ni compétence poussée en administration. Au travers de ces travaux, l'IFB souhaite non seulement mettre à disposition des ressources de calcul mais également permettre à toute organisation le souhaitant d'adopter et enrichir ses procédures de gestion et de déploiement afin de faciliter la mise en place d'une infrastructure de calcul clé en main.

WAVES: a Web Application for Versatile Enhanced bioinformatic Services

Marc CHAKIACHVILI[1,2], Sylvain MILANESI[1], Anne-Muriel CHIFOLLEAU[1] and Vincent LEFORT[1]

[1] LIRMM, CNRS & Univ. Montpellier, Montpellier, France

[2] European Molecular Biology Laboratory, European Bioinformatics Institute, Wellcome Genome Campus, Hinxton, Cambridge, United Kingdom

Corresponding Author: vincent.lefort@lirmm.fr

1 Introduction

Any new bioinformatic tool must be made available to its user's community, essentially to biologists, for whom command line interfaces are often cumbersome. This need is mainly satisfied by implementing a dedicated website. Several solutions, such as Galaxy [1] or Mobyle [2], were developed to ease tool integration within automatically-generated web pages, making them accessible through a generic web user interface. These generic approaches allow the integration of a large variety of bioinformatic tools behind the same interface model. However, these interfaces are generally poorly customizable, preventing web developers from creating high-level and interactive services adapted to each scientific community needs.

Here, we present a versatile service-oriented web application, named WAVES, designed to provide an integrated web-oriented interface for bioinformatic tools, as a facade [3] that conceals the complexity of the underlying computing architecture. The main goal of WAVES is to gather a comprehensive selection of bioinformatic services within a single application programming interface (API). It may integrate tools from different environments and remote resources. In this way, WAVES allows bioinformaticians to integrate tools easily so they can focus on designing high-level user interfaces for community specific web applications.

2 Features

There are three different ways to interact with WAVES services: web pages, web forms, and a RESTful API. WAVES automatically creates a web page for each integrated tool. This basic feature is essential for providing end-users with an interface that enables them to run online bioinformatic analyses. In the same manner, it generates web forms to be directly integrated into any website. Lastly, WAVES provides web service entries in its RESTful API, thus generating services suitable for software interoperability. These web services all share the same API structure which complies with Core API [4].

WAVES is compatible with a variety of computing infrastructures. It runs any locally installed tool. By setting the required credentials, WAVES runs remotely installed tools through a secure network connection. It interoperates with most computing resource management systems. For interoperability purposes, WAVES interacts with Galaxy. It lists the tools available in Galaxy instances and offers the ability to import them automatically as new services. WAVES can then run the tools within the Galaxy instance from which it was imported, check computation status, and retrieve results.

Funding

This work was supported by Institut Français de Bioinformatique [ANR-11-INBS-0013].

References

[1] Goecks, J. et al. Galaxy: a comprehensive approach for supporting accessible, reproducible, and transparent computational research in the life sciences. *Genome Biology*, 11(8), R86, 2010.

[2] Neron, B. et al. Mobyle: a new full web bioinformatics framework. *Bioinformatics*, 25(22), 3005–3011, 2009.

[3] Gamma, E., ed. Design Patterns: Elements of Reusable Object-Oriented Software. *Addison-Wesley Professional Computing Series*. Addison-Wesley, Boston, MA, USA, 1995.

[4] http://www.coreapi.org/.

Posters

A clinical bioinformatics framework for single-cell profiling of rare diseases

Loredana Martignetti[1,§], Akira Cortal[1], Steicy Sobrino[3,4], Emmanuelle Six[3,4], Marina Cavazzana[3,4,5,6], Antonio Rausell [1,2,3 *]

[1] Clinical Bioinformatics Laboratory, Imagine Institute, Paris , France
[2] INSERM UMR 1163, Institut Imagine,Paris, France
[3] Paris Descartes University, Sorbonne Paris Cité Paris, France
[4] Laboratory of Human Lymphohematopoiesis, INSERM UMR 1163, Imagine Institute, Paris, France
[5] Biotherapy Department, Necker Children's Hospital, Assistance Publique-Hôpitaux de Paris, Paris
[6] Biotherapy Clinical Investigation Center, Groupe Hospitalier Universitaire Ouest, Assistance Publique-Hôpitaux de Paris, INSERM, Paris

(§) Presenting author
(*) Corresponding Author: antonio.rausell@inserm.fr

1. Introduction

Out of more than 4,000 Mendelian diseases clinically described to date, around 50% still lack the identification of their causal gene or variant [1]. Transcriptional profiling of the affected organs and tissues may contribute to the characterization of the molecular and cellular causes of a disease. Cell heterogeneity uncovered through single-cell RNA-seq may identify the relevant cell types or cell states responsible of the onset and progression of these diseases. Notwithstanding, comparative analyses across patients and control samples are challenged by (i) technical stochasticity and batch effects, and (ii) environmental and physiological factors including age, sex, life style and clinical history [2].

2. Results

Here we present a comprehensive bioinformatics framework for the single-cell transcriptional analysis of rare genetic diseases developed by the Clinical Bioinformatics Lab at the *Imagine* Institute. First, Cell-ID, a method based on Multiple Correspondence Analysis, is applied to extract a *cell identity card* in the form of an unbiased per-cell gene signature for each individual cell in a dataset. Per-cell signatures, or *cell fingerprints*, allow: (i) automatic cell type prediction using reference cell type signatures, and (ii) functional enrichment analysis using gene sets representing functional ontologies and pathways. More interestingly for the study of rare diseases, Cell-ID is able to identify rare sub-population of cells within a sample (<2%), and subsequently to "*blast*" such cell signatures against reference datasets, i.e.: to test for a statistically robust replication of the newly uncovered cell signatures across samples from different patients. In benchmark datasets, the method was able to overcome batch effects associated to different donors, tissues of origin, and sequencing technologies. A complementary method of Cell-ID is Sample-ID, based on Single Value Decomposition, is developed to extract a *sample identity card* in the form of unbiased gene signatures characterizing the observed transcriptional heterogeneity within a sample. Per-sample signatures, or *sample fingerprints*, allow in turn to "*blast*" *query* patient samples against reference single-cell RNA-seq libraries from genetically-characterized patients, thus favoring molecular diagnosis. Both Cell-ID and Sample-ID are being systematically applied to single-cell RNA-seq datasets from (i) the Human Cell Atlas project, profiling healthy human organs and tissues, and (ii) in-house collections of rare disease patients profiling the affected organs and tissues. Pre-computed per-cell and per-sample gene signatures are being systematically generated to provide the community with a reference library of healthy and rare disease transcriptional hallmarks at single-cell level. Such hallmarks may ultimately translate into molecular biomarkers with a clinical diagnostic value.

References

1. Chong JX, Buckingham KJ, Jhangiani SN, Boehm C, Sobreira N, Smith JD, et al. The genetic basis of Mendelian phenotypes: discoveries, challenges, and opportunities. Am J Hum Genet. 2015;97(2):199–215
2. Oliver Stegle, Sarah A Teichmann and John C Marioni. Computational and analytical challenges in single-cell transcriptomics. *Nat Rev Genet.* 2015 Mar;16(3):133-45.

A graph theoretical approach to depicting sex-biased dispersal in ancient populations: mitochondrial DNA *vs.* Y-chromosome variation

Pierre Justeau[1], Simão Moreira Rodrigues[2], Maria Pala[3], Ceiridwen Edwards[4], Martin B. Richards[5]

Archaeogenetics Research Group, Department of Biological and Geographical Sciences, School of Applied Sciences, University of Huddersfield, Huddersfield, HD1 3DH, UK

Corresponding author: pierre.justeau@hud.ac.uk

Keys words: ancient DNA, graph theory, mitochondrial DNA, Y chromosome, sex-biased dispersal

Since the advent of next-generation sequencing, the quality and quantity of ancient DNA data have allowed us to depict human movements through time and more precisely the contribution of different sexes to these dispersals [1,2,3]. One of the more surprising examples is the male migration during the Bronze Age from the Eurasian steppe to western Europe. This sex-biased dispersal led to an important genetic turnover by virtually replacing the previous Neolithic Y-chromosome diversity with haplogroup R1b-M269 [2].

However, these studies do not give us precise information about the genetic diversity of the haploid marker systems (mitochondrial DNA and Y-chromosome haplogroups) and the social structure of the R1b-M269 group. Thus, we have developed, using graph theory, an approach that considers a more precise comparison of genetic and archaeological features to better understand the processes and consequences of this important episode in Eurasian history.

[1] Silva Marina, Oliveira Marisa, Vieira Daniel, Brandão Andreia, Rito Teresa, Pereira Joana, Fraser Ross, Hudson Bob, Gandini Francesca, Edwards Ceiridwen, Pala Maria, Koch John, Wilson James, Pereira Luísa, Richards Martin, Soares Pedro. A genetic chronology for the Indian Subcontinent points to heavily sex-biased dispersals. *BMC Evolutionary Biology*. 17,2017.

[2] Olalde Iñigo, Armit Ian, Kristiansen Kristian, Pinhasi Ron, Haak Wolfgang, Barnes Ian, Lalueza-Fox Carles , Reich David, Bonsall Clive, Brace Selina, E. Allentoft Morten, Booth Thomas, Rohland Nadin, Mallick Subhashis, Szécsényi-Nagy Anna, Mittnik Alissa, Altena Eveline, Lipson Mark, Lazaridis Iosif, Krause Johannes et al. The Beaker Phenomenon and the Genomic Transformation of Northwest Europe. *Nature*. 555, 2018

[3] Olalde Iñigo, Mallick Subhashis, Patterson Nick, Rohland Nadin, Villalba-Mouco Vanessa, Silva Marina, Dulias Katharina, J. Edwards Ceiridwen, Gandini Francesca, Pala Maria, Soares Pedro, Ferrando-Bernal Manuel, Adamski Nicole, Broomandkhoshbacht Nasreen, Cheronet Olivia, J. Culleton Brendan, Fernandes Daniel, Marie Lawson Ann, Mah Matthew et al. The genomic history of the Iberian Peninsula over the past 8000 years. *Science*. 363, 2019

A novel DNA methylation signature for cell-type deconvolution in immuno-oncology

Ting XIE and Vera PANCALDI

Centre de recherche en Cancérologie de Toulouse (CRCT, UMR1037 Inserm / Université Toulouse III Paul Sabatier, 2 Avenue Hubert Curien, 31037, Toulouse, France)

Corresponding Author: ting.xie@inserm.fr, vera.pancaldi@inserm.fr

1. Introduction

Much of the recent progress in cancer treatment derives from the exploitation and reactivation of immune cells that are infiltrating the tumour micro-environment. Despite the great potential of immuno-oncology, there is a great difference in efficacy of these therapies across tumour-types and patients. It is thus of paramount importance to develop tools to identify the different types of immune cells present in biopsy samples. DNA methylation profiles are cell-type specific and an excellent alternative to transcriptomes to perform cell-type deconvolution [1,2]. So far, several reference-based deconvolution methods based on DNA methylation have been proposed [1]. As for deconvolution based on RNAseq, the procedure usually involves constructing a signature which is specific to the problem of interest. So far, available methods have only used signatures based on Illumina human 450k or 850k array and are usually limited to the most widely studied immune cells types from blood, such as T cells (CD4+, CD8+), neutrophils, B cells, NK cells, and monocytes.

2. Results

In this project, we aimed to deconvolute the presence of myeloid cells in tumour samples, to study the role that they play in repressing lymphocytes in their normal anti-tumour functions or in altering the effectiveness of immunotherapies. We exploited a large collection of haematopoietic epigenomes [3], to establish a novel DNA methylation signature based on whole-genome bisulfite sequencing (WGBS). We selected CpGs from the WGBS data which overlap the 850k array probes and tested the resulting signature with the EpiDISH package using the robust partial correlation method [4]. Our newly generated DNA methylation signature is able to distinguish 12 cell types, including different types of macrophages that can be inflammatory or activated (macrophages M0, M1 and M2), based on 107 samples from purified cells. We annotated the new signature and tested the performance on public datasets featuring both DNA methylation and flow cytometry estimations of cell type proportions in peripheral blood mononuclear cells, finding a good correlation. To test its performance on tumour samples, we applied our signature to 32 Lung Adenocarcinoma samples from TCGA (TCGA-LUAD).

The new signature and the developed pipeline will be used in the future to perform cell-type deconvolution of tumour samples, as a first step towards gaining a better understanding of the composition of the tumour micro-environment in different cancers.

References

1. Eugene A Houseman et al. DNA Methylation Arrays as Surrogate Measures of Cell Mixture Distribution. *BMC Bioinformatics* 13(86), 2012.
2. Winston Timp et al. Large Hypomethylated Blocks as a Universal Defining Epigenetic Alteration in Human Solid Tumors. *Genome Medicine* 6(8), 2014.
3. M. Farlik et al. DNA methylation dynamics of human hematopoietic stem cell differentiation, Cell Stem Cell 19(6): 808-22, 2016.
4. Andrew Teschendorff, et al. A comparison of reference-based algorithms for correcting cell-type heterogeneity in Epigenome-Wide Association Studies. *BMC Bioinformatics* 18 (1):105, 2017.

A set of methods to study three classes of non-coding RNAs

Maxime Delmas[1], Lisa Muniz[1], Didier Trouche[1], Estelle Nicolas[1], Marion Aguirrebengoa[1]

[1] LBCMCP, Université Paul Sabatier *118 Route de Narbonne*, Toulouse, France

Corresponding Author: maxime.delmas@etu.univ-rouen.fr

Non-coding RNAs (ncRNAs) are a very large class of RNAs whose involvement in cellular processes has been underestimated for a long time. Their heterogeneity in term of biosynthesis, structure and localization makes them difficult to detect by methods used for their protein-coding analogues.
We are more particularly interested in three classes of unannotated ncRNAs: circular RNAs [1] formed by back-splicing, vlinc RNAs [2] (very long intergenic ncRNAs of minimum 50 kb), and read-through RNAs formed by defects in gene transcription termination [3]. We developed a set of methods to identify these different classes of ncRNAs.
For circular RNAs detection, our developed pipeline uses fragment reconstruction from paired-end data to optimize the detection of chimeric reads from which circular reads are detected. The pipeline combines two chimera detection tools with three circular detection tools. A filter was developed to eliminate chimeric reads from STAR that produce circular RNA artifacts. A complete annotation step, to categorize and describe circular RNAs, and a linked Shiny application, to carry out further analyses such as differential expression, were implemented in the pipeline. Our pipeline detection performance was compared to two other pipelines (CirCompara [4] and circtools [5]) in a benchmark step using a Control versus RNAse R-treated RNA-seq dataset (RNAse R digests linear RNAs and allows the enrichment of circular RNA reads).
The vlinc RNAs detection method is based on the aggregation of coverage windows in intergenic regions respecting a maximum gap length to predict vlinc RNAs coordinates [6], which initially used the IGB interface. We recoded this method and improved the accuracy of the predicted coordinates. This method was also adapted to detect read-through RNAs. However, this approach requires several hyperparameters, which are difficult to set up and justify. So, in the read-through RNA context, we developed a new detection method based on their modelization using HMM (Hidden Markov Model) from the coverage in intergenic regions and in introns of the upstream gene. This method seems to be more efficient and accurate than the aggregation method and should be generalized for vlincRNAs studies. Preliminary results from RNA-Seq datasets of two biological conditions (cellular proliferation versus senescence) are presented for each described methods.

References

[1] Salzman, J., Gawad, C., Wang, P. L., Lacayo, N. & Brown, P. O. Circular RNAs Are the Predominant Transcript Isoform from Hundreds of Human Genes in Diverse Cell Types. PLoS ONE 7, e30733 (2012).
[2] Laurent, G. S. et al. Functional annotation of the vlinc class of non-coding RNAs using systems biology approach. Nucleic Acids Res. 44, 3233–3252 (2016).
[3] Muniz, L. et al. Control of Gene Expression in Senescence through Transcriptional Read-Through of Convergent Protein-Coding Genes. Cell Rep. 21, 2433–2446 (2017).
[4] Gaffo, E. et al. CirComPara: A Multi‐Method Comparative Bioinformatics Pipeline to Detect and Study circRNAs from RNA‐seq Data. Non-Coding RNA *3*, 8 (2017).
[5] Jakobi, T et al. circtools - a one-stop software solution for circular RNA research. Bioinformatics (2018).
[6] Kapranov, P. et al. The majority of total nuclear-encoded non-ribosomal RNA in a human cell is "dark matter" un-annotated RNA. BMC Biology *8*, 149 (2010).

A state-of-the-art analysis of innovation software tools for primary analysis for Oxford Nanopore sequence data

Charlotte BERTHELIER[1], Bérengère LAFFAY[1,2], Sophie LEMOINE[1] and Laurent JOURDREN[1]

[1] École normale supérieure, PSL Research University, CNRS, Inserm, Institut de Biologie de l'École normale supérieure (IBENS), Plateforme Génomique, 75005 Paris, France

[2] Master Bioinformatique, Normandie Université, UNIROUEN

Corresponding Author: charlotte.berthelier@biologie.ens.fr

Keywords: Oxford Nanopore Technologies, direct RNA sequencing, MinION, basecalling, Albacore, Guppy, MinKNOW and MinIT.

Summary

The Oxford Nanopore Technologies (ONT) sequencing is in constant evolution. Progress is emerging, especially for the primary analysis of full-length total DNA-and RNA-sequencing (also called direct RNA sequencing), which are provided by the ONT sequencers (MinION, GridION and PromethION) [1].

Instead of capturing images like in Illumina sequencing, the MinION sequencer captures electronic signals. Basecalling, the computational process of translating raw electrical signal to nucleotide sequence, have a critical importance to generate high-quality data. With the recent updates of the ONT basecaller programs, this poster examine the performance of two different basecallers: Albacore and Guppy, using cDNA on R9.4 flowcells with the 1D chemistry.

Our poster includes a ToulligQC [2] quality control reports comparison of the two basecallers results. In addition, it will review other innovations such as the new release of the Oxford Nanopore's MinKNOW acquisition software and the incoming of the MinIT device [3]. In addition, we also benchmarked many factors that influence basecalling speed such as software version, HD vs SSD, thread number, CPU vs GPU…

The main goal of this study is to reduce the computation time required by basecalling and QC control steps to provide as soon as possible data to analyze to our users once runs has finished.

References

[1] Miten Jain et al. The Oxford Nanopore MinION: delivery of nanopore sequencing to the genomics community. Genome Biol., 17(1): 239, 2016.

[2] https://github.com/GenomicParisCentre/toulligQC

[3] https://nanoporetech.com/products/minit

A tool for very fast taxonomic comparison of genomic sequences

Martial BRIAND[1], Mariam BOUZID[1], Marc LEGEAY[1,2], Marion FISCHER-LE SAUX[1], Claire LEMAITRE[3], Gilles HUNAULT[4] and Matthieu BARRET[1]

[1] IRHS-INRA, 42, rue Georges Morel, 49071, Beaucouzé, France
[2] LERIA, 2 bd Lavoisier, 49045, Angers, France
[3] GenScale INRIA IRISA, Campus de Beaulieu, 35042, Rennes, France
[4] HIFI, 4, Rue Larrey, 49933, Angers, France

Corresponding Author: martial.briand@inra.fr

En taxonomie bactérienne, la mesure des identités nucléotidiques moyennes (ANI) [1] est maintenant largement utilisée pour regrouper les génomes bactériens en un ensemble phylogénétique ou clique. Cependant, le temps de calcul de cette distance, basée sur des recherches d'homologies par blast, peut être long et limitant pour l'analyse de grands jeux de données. A titre d'exemple, 194140 génomes prokaryotes sont disponibles au 21 mars 2019 au NCBI.

Nous proposons ici d'utiliser le pourcentage de K-mers partagés entre les génomes pour estimer leur proximité phylogénétique. Nous comparons cette distance (calculée avec Simka [2]) avec la valeur d'ANI (calculée avec pyani [3]) sur un jeu de 944 génomes de Pseudomonas spp. publiquement disponibles, ainsi que les temps de calcul de ces matrices.

Les arbres générés à partir de grandes matrices de distances étant parfois difficilement lisible, nous avons développé des représentations graphiques originales de ces matrices. L'outil "Ki-S" que nous proposons, permet de générer les matrices de distances ANIb et k-mers dans un environnement Galaxy et fournit des représentations originales pour améliorer la visualisation et l'analyse des grandes matrices de similarité.

References

1. Richter, M., Rossello-Mora, R.: Shifting the genomic gold standard for the prokaryotic species definition. *Proc Natl Acad Sci* USA 106: 19126-19131. DOI :10.1073/pnas.0906412106, 2009.
2. Benoit G, Peterlongo P, Mariadassou M, Drezen E, Schbath S, Lavenier D, Lemaitre C. : Multiple comparative metagenomics using multiset k-mer counting. PeerJ Computer Science 2:e94. DOI: 10.7717/peerj-cs.94, 2016.
3. Pritchard Leighton and Glover, Rachel H. and Humphris, Sonia and Elphinstone, John G. and Toth, Ian K: Genomics and taxonomy in diagnostics for food security: soft-rotting enterobacterial plant pathogens. *Anal. Methods,* 2016, 8, 12-24. DOI: 10.1039/C5AY02550H, 2016.

A web server for identification and analysis of coevolution in overlapping proteins

Elin Teppa[1], Alessandra Carbone[1,2]

1 Laboratory of Computational and Quantitative Biology (LCQB), Sorbonne Université, CNRS, IBPS, UMR7238, 4, place Jussieu 75005 Paris, France.
2 Institut Universitaire de France (IUF) 75005 Paris, France

Corresponding author:elinteppa@gmail.com

Abstract

Overlapping genes exist in all domains of life and are especially abundant in viral genomes. The existence of overlapping reading frames increases the rising of deleterious mutations for one of the proteins, since a single nucleotide substitution may affect both proteins. Molecular coevolution may be seen as a mechanism to tolerate or compensate unfavorable mutations, decreasing the evolutionary constraints in the overlapping region. For instance, a favorable mutation in one reading frame may be unfavorable in the other reading frame and additional mutations may be needed to compensate the first mutation. Although molecular coevolution was widely used in viral genomes, the "overlap problem" was disregarded. Here, we present a server that facilitates the analysis of coevolution in overlapping proteins and of the impact of mutations in another ORF.

Keywords: coevolution; compensatory mutations, virus, overlapping proteins

Introduction

Multiple studies of coevolving positions in viral sequences have been useful to understand functionally significant residues [1,2], to predict protein-protein interactions [3], to modulate viral fusion [4] and to identify drug resistance mutations [5–9] among others.

The genomes of most viral species have overlapping genes—two or more proteins coded for by the same nucleotide sequence. ORFs may overlap in various manners considering the type, the direction of transcription and the ORFs' phase (Fig 1). Sequence analysis in overlapping ORFs represents a challenge due to changes in the nucleotide sequence that may simultaneously affect both proteins within their overlapping region.

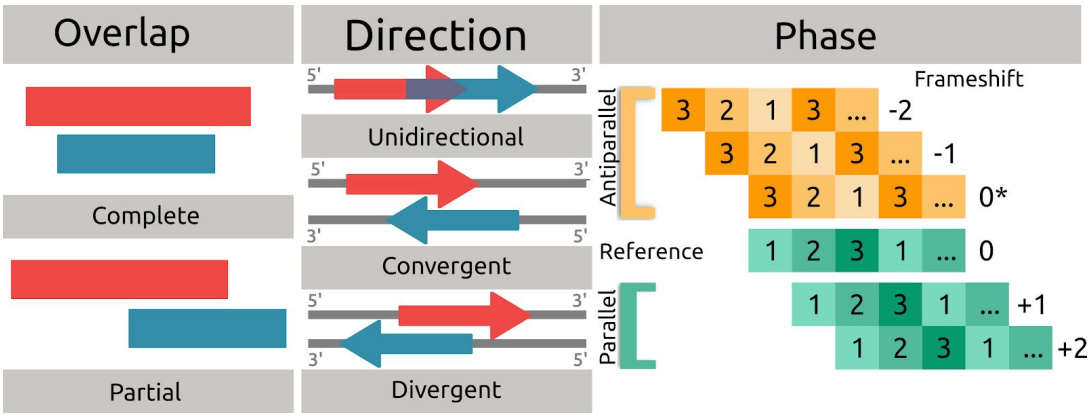


Figure 1: Definitions on ORFs overlap.

An overlap between two ORFs can be complete (if an ORF is nested within the other) or partial (if only the 3' or 5' end are overlapping). ORFs can overlap on the same strand, or in the case of a double-stranded genome, on the reverse complementary strand. Hence, three directions are possibles: unidirectional, convergent and divergent. The reference ORF, in a pair of overlapping ORFs, is called phase 0. Overlaps in a parallel strand can be in two phases whereas antiparallel-strand overlaps can be in three phases.

Given that coevolution may be seen as a mechanism to tolerate or compensate unfavorable mutations, molecular coevolution in the overlapping region may help to decrease evolutionary constraints. As far as we know, there is no study of coevolution that considers both overlapped proteins.

In the overlapping region, coevolution in an ORF: may be mirrored by coevolution in the other ORF; may generate a non-synonymous substitution which in turn may be compensated by other mutations (inside or outside the overlapping region); may generate synonymous substitutions (Fig 2).

The motivation for this server is to provide a tool to facilitate the analysis of coevolution in overlapped protein and of the impact of mutations in another ORF. To do that we combine information at protein and nucleotide levels.

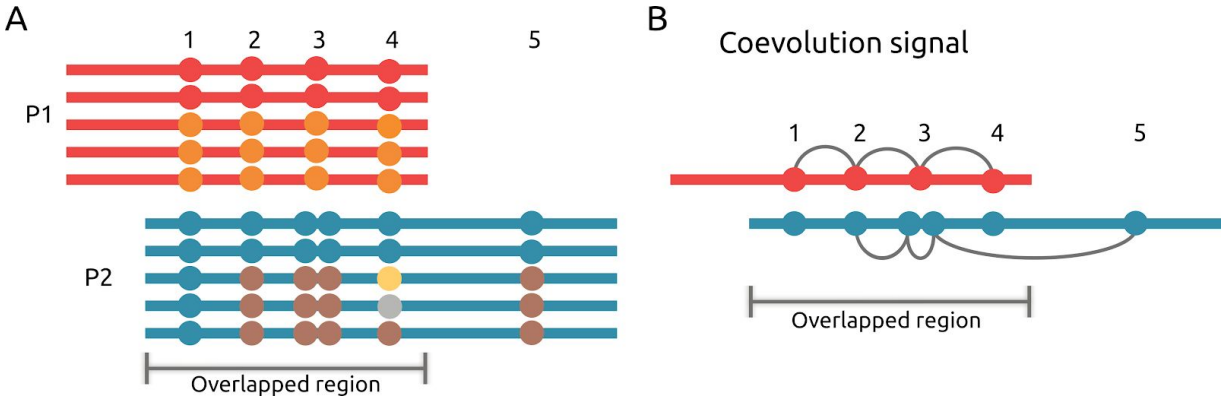


Figure 2: Coevolution pattern in overlapping region.

Different effects of four coevolving positions in the overlapped region of two proteins (P1 and P2). **A:** A cluster of four coevolving positions is represented in P1's alignment where two sequences maintain the wild-type residues (red circles) and three sequences show mutations on all positions (orange circles). A

mutation in P1 may be coupled by synonymous substitutions in P2 (column 1); the same non-synonymous substitution (column 2), two non-synonymous substitutions in adjacent positions (column 3); a variety of non-synonymous substitutions (column 4). The cluster of coevolving positions may also contain positions outside the overlapping region (column 5). **B:** P1 shows a coevolution signal between the first four positions (gray lines) which partially coincides with coevolution detected in P2.

Methods

Input

The input is a nucleotide alignment of the pair of overlapping protein sequences to be analyzed. It will contain the overlapped and non-overlapped regions of both proteins, as well as the start and end positions of the proteins and their corresponding DNA strand (parallel or antiparallel) (Figure 1).

Workflow

Given a DNA alignment and its associated distance tree that can be provided or optionally generated automatically, all subsets of sequences corresponding to the subtrees of the tree are systematically considered for coevolution analysis. For each subset, the ORF1 and ORF2 sequences (Fig 2) are translated into amino acids and the resulting protein alignments are used as input to predict coevolving positions using the BIS2 algorithm [10,11]. Our iterative strategy allows applying BIS2 in a large number of conserved sequences. As part of the result, the clusters of coevolving positions detected for both proteins are provided. If coevolution is detected in the overlapping region for one of the proteins, the effect of variation is analyzed in the other protein. By analyzing the subset of sequences where the cluster is detected for the first protein, we identify if the coevolving positions are accompanied by one or more synonymous/non-synonymous substitution(s) and if these positions also show coevolution in the second protein.

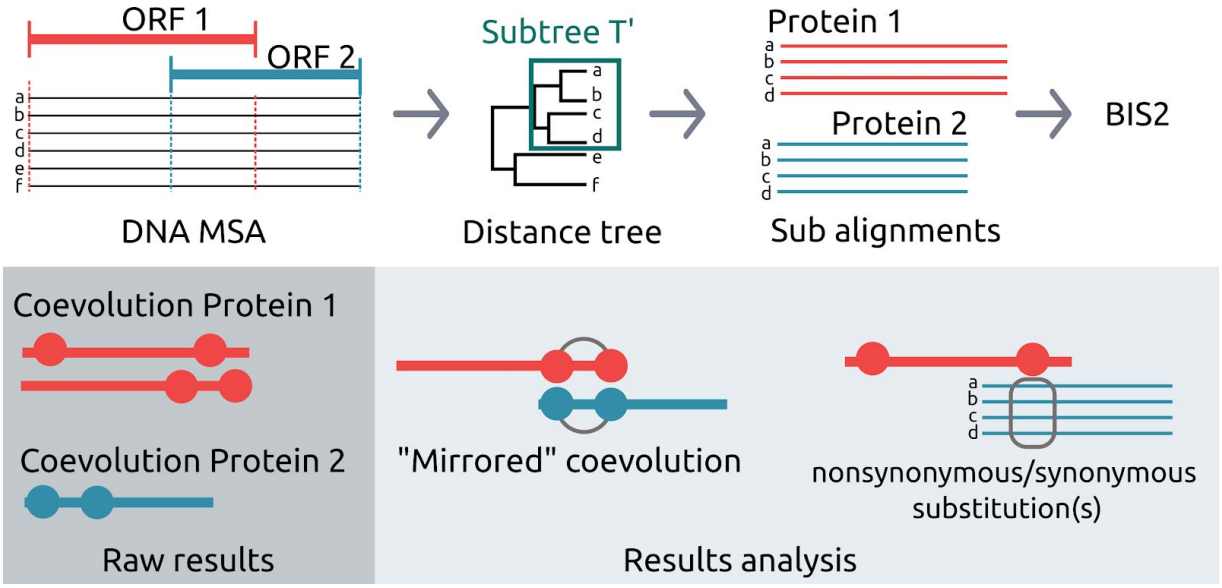


Figure 3: Schematic representation of the workflow from the input sequences to the results.

The DNA alignment covering both ORFs to be analyzed is used to generate a distance tree, optionally the tree may be provided by the user. Then, the tree is partitioned in all possible subtrees. The protein sequences corresponding to the subtrees are used as input to compute coevolution using BIS2 algorithm. The results include the coevolution of each of the proteins, as well as the effect of the mutations of one protein on the other. It is also indicated if both proteins show coevolution in equivalent positions ("mirrored" coevolution) or if the mutation of the co-evolved position in a protein is accompanied by synonymous or nonsynonymous mutations in the other.

Conclusions

We have developed an interactive web server providing an intuitive representation of of the coevolved residues predicted in overlapping proteins. To the best of our knowledge, this is the only publicly available method designed to analyze coevolution in overlapping protein sequences. The server is simple to use and it provides a powerful tool the virologist and the biologist to compute coevolution and analyze the effect of mutations in overlapping regions. Its results should help to elucidate the evolutionary constraints found in overlapping ORFs.

Acknowledgements

This work was supported by the French "Agence Nationale de la Recherche sur le SIDA et les hépatites virales" (ANRS CSS4 ECTZ25224 – 2017-19 to AC; www.anrs.fr).

References

1. Le L, Leluk J. Study on phylogenetic relationships, variability, and correlated mutations in M2 proteins of influenza virus A. PLoS One. 2011;6: e22970.
2. Jain J, Mathur K, Shrinet J, Bhatnagar RK, Sunil S. Analysis of coevolution in nonstructural proteins of chikungunya virus. Virol J. 2016;13: 86.
3. Champeimont R, Laine E, Hu S-W, Penin F, Carbone A. Coevolution analysis of Hepatitis C virus genome to identify the structural and functional dependency network of viral proteins. Sci Rep. 2016;6: 26401.
4. Douam F, Fusil F, Enguehard M, Dib L, Nadalin F, Schwaller L, et al. A protein coevolution method uncovers critical features of the Hepatitis C Virus fusion mechanism. PLoS Pathog. 2018;14: e1006908.
5. Rhee S-Y, Liu TF, Holmes SP, Shafer RW. HIV-1 subtype B protease and reverse transcriptase amino acid covariation. PLoS Comput Biol. 2007;3: e87.
6. Handel A, Regoes RR, Antia R. The role of compensatory mutations in the emergence of drug resistance. PLoS Comput Biol. 2006;2: e137.
7. González-Ortega E, Ballana E, Badia R, Clotet B, Esté JA. Compensatory mutations rescue the virus replicative capacity of VIRIP-resistant HIV-1. Antiviral Res. 2011;92: 479–483.
8. Bloom JD, Gong LI, Baltimore D. Permissive secondary mutations enable the evolution of influenza oseltamivir resistance. Science. 2010;328: 1272–1275.
9. Tanaka MM, Valckenborgh F. Escaping an evolutionary lobster trap: drug resistance and compensatory mutation in a fluctuating environment. Evolution. 2011;65: 1376–1387.
10. Dib L, Carbone A. Protein fragments: functional and structural roles of their coevolution networks. PLoS One. 2012;7: e48124.
11. Oteri F, Nadalin F, Champeimont R, Carbone A. BIS2Analyzer: a server for co-evolution analysis of conserved protein families. Nucleic Acids Res. 2017;45: W307–W314.

A workflow based on self-organizing map for clustering stable structures of proteins from molecular dynamics simulations

Philippe Noel, Emmanuel Bresso, Dave Ritchie, Bernard Maigret and Marie-Dominique Devignes
Université de Lorraine, CNRS, Inria, LORIA, F-5400 Nancy

Corresponding author: philippe.noel@inria.fr

Self-organizing map [1] is a clustering method that maps high-dimensional data onto a two-dimensional grid such that similar objects are placed close to each other. It has been proposed as an alternative method to classical RMSD (Root Mean Square Deviation) heatmaps for clustering frames after a molecular dynamics (MD) simulation [2]. However no workflow dedicated for this task is yet available to run and test the method.

Our workflow (SOM4MD) is composed of three parts. (1) The frame preparation takes as input the dcd files resulting from the molecular dynamics run (using NAMD) and prepares the data for SOM execution. (2) The SOM analysis is distributed on a cluster on the MBI platform* to speed up execution. (3) Clustering results are processed for automatic selection of representative frames from the most compact clusters, followed by mapping on a standard RMSD heatmap for visual inspection and validation.

The SOM4MD workflow has been tested on MD simulations of $1\mu s$ to $100\mu s$ with up to two million frames. Time-length of execution depends on the number of input frames and of CPU used. Automatic selection of representative frame selection is satisfying so far from expert point of view but requires more case-studies. Nevertheless this method provides significant time saving and an objective basis in MD simulations interpretation. The workflow is programmed in Python and is available on demand for more testing and feedbacks.

References

[1] Teuvo Kohonen. Self-organized formation of topologically correct feature maps. *Biological Cybernetics*, 43(1):59–69, 1982.

[2] Guillaume Bouvier, Nathan Desdouits, Mathias Ferber, Arnaud Blondel, and Michael Nilges. An automatic tool to analyze and cluster macromolecular conformations based on self-organizing maps. *Bioinformatics*, 31(9):1490–1492, May 2015.

A workflow to analyse single-cell transcriptomes from heterogeneous tumors

Léa BELLENGER [1], Mirca S. SAURTY [2], Ghislaine MORVAN-DUBOIS [2], Hervé CHNEIWEISS [2], Marie-Pierre JUNIER [2], Christophe ANTONIEWSKI [1]

[1] ARTbio bioinformatic platform, 9 Quai St Bernard, 75005, Paris, France

[2] UMR 8246 - Neuroscience and Glial plasticity, 7 Quai St Bernard, 75005, Paris, France

Corresponding Author: lea.bellenger@sorbonne-universite.fr

Over the past decade, single-cell sequencing has revolutionized transcriptomic analysis. It is now possible to capture gene expression at a cell level and thereby to better understand complex multicellular processes. Its potential impact made the scRNAseq analysis as one of our priority areas.

In collaboration with the Glial plasticity and Neuro-oncology team at the IBPS, we developed a new data reduction approach to analyse heterogeneous brain tumors at single-cell resolution. In brief, we computed a tumorigenic score for each Glioblastoma single-cell transcriptome according to the expression of a signature set of genes. Cells were then split into two groups of "high tumorigenicity" and "low tumorigenicity" cells and differential expression analysis between these two groups was performed using Mann-Whitney tests. The approach was applied to single-cell RNAseq from 4 glioblastomas [1] and the TCGA collection of bulk transcriptomes from 155 glioblastomas. Despite their very distinct origins (single-cell versus tissue RNAseq and 4 vs 155 tumors), we found a remarkable overlap between differentially expressed genes in these two series of datasets, thereby validating our data reduction approach.

With the aim to provide transparent and reproducible computational methods, we coded a set of functionally independent R scripts and linked these scripts in a workflow that allows to easily adapt our approach to any single-cell RNAseq data.

References

1. Darmanis S, et al; Single-Cell RNA-Seq Analysis of Infiltrating Neoplastic Cells at the Migrating Front of Human Glioblastoma. *Cell Rep* Oct 31;21(5):1399-1410, 2017.

A workflow to build a relevant bacterial genome sub-dataset from public databases

Sam AH-LONE[1], Sandra DÉROZIER[1], Valentin LOUX[1] and Hélène CHIAPELLO[1]

[1] MaIAGE, INRA, Université Paris-Saclay, 78350, Jouy-en-Josas, France

Corresponding Author: sam.ah-lone@inra.fr

Thanks to rapid progress in High-Throughput Sequencing (HTS) technologies, more than 190,000 bacterial assemblies are now available in public databases [1]. For a few bacterial species of interest, more than 10,000 strains have been sequenced, some of them with very similar genomic content while exhibiting heterogeneous assembly quality levels. Unfortunately, most genome comparison tools are not yet scalable to those large datasets. In this work we propose an accessible and scalable tool to rapidly analyze and filter large sets of closely related bacterial genomes.

We set up rules to build a representative sub-dataset by taking into account assembly quality while maintaining the genomic diversity of the original dataset. For this, assemblies are first compared and clustered using Mash distance [2] and the Neighbor-Joining algorithm. Then most reliable representatives of each cluster are chosen using assembly quality metrics, such as N50, contig numbers, and assembly lengths computed with Quast [3].

We designed a Snakemake [4] pipeline to download, analyse and filter a bacterial assembly dataset from RefSeq using the defined rules. The procedure has been first tested on two datasets of 300 assemblies from *S.enterica* and *B.subtilis*. It is currently being evaluated on 9,520 *S.enterica* chromosome assemblies.

References

[1] NCBI Genbank FTP site. ‘prokaryotes.txt’. [en ligne]. https://ftp.ncbi.nlm.nih.gov/genomes/GENOME_REPORTS/prokaryotes.txt (14/03/2019).

[2] Ondov BD, Treangen TJ, Melsted P, Mallonee AB, Bergman NH, Koren S, et al. Mash: fast genome and metagenome distance estimation using MinHash. Genome Biol. 2016;17:132.

[3] Gurevich A, Saveliev V, Vyahhi N, Tesler G. QUAST: quality assessment tool for genome assemblies. Bioinformatics. 2013;29:1072–5.

[4] Köster, Johannes and Rahmann, Sven. Snakemake - A scalable bioinformatics workflow engine. Bioinformatics 2012.Gurevich A, Saveliev V, Vyahhi N, Tesler G. QUAST: quality assessment tool for genome assemblies. Bioinformatics. 2013;29:1072–5.

Visualize Advanced Data Comparisons with BiocompR.

Yoann Pageaud[1], Pavlo Lutsik[1]

[1] Computational Cancer Epigenomics – DKFZ, Im Neuenheimer Feld 280, 69120, Heidelberg, Germany

Corresponding Author: yoann.pageaud@gmail.com

Increasing amount of omics data generated since the last decade has led to the necessity of finding standardized ways of analysing and representing them. The R programming language has quickly established itself as the reference for statistical analysis and graphics on omics data, jump-started by the Tidyverse packages[1], among which Ggplot2[2] has brought to scientists a new grammar[3] for plotting graphs using combinations of independent components. Here I introduce *BiocompR*, a R Package that uses Ggplot2, to update some often used plots dedicated to data comparisons, dataset exploration and, ultimately, to provide the user with versatile and customizable graphics. In the near future, BiocompR will be utilized by the Methrix[4] package – which is being developed by our group – to visualize genome-wide methylation/coverage information from whole genome bisulfite sequencing datasets.

References

[1] Wickham H. *Tidyverse: Easily Install and Load 'Tidyverse' Packages.* package version 1.2.1, 2017.
[2] Wickham H. *Ggplot2: Elegant Graphics for Data Analysis.* Springer, 2009.
[3] Leland W. *The Grammar of Graphics.* Springer, coll. "Statistics and Computing", 2005, 2e ed.
[4] Methrix, an R Package - https://github.com/CompEpigen/methrix

ALFA: Annotation Landscape For Aligned reads

Mathieu BAHIN[1], Benoit F Noël[1,2], Valentine Murigneux[3], Charles Bernard[1], Leila Bastianelli[1,3], Hervé Le Hir[3], Alice Lebreton[2] and Auguste GENOVESIO[1]

[1] Computational Bioimaging and Bioinformatics Team
[2] Bacterial Infection and RNA Destiny Team
[3] Expression of eukaryotic messenger RNAs Team
IBENS, Département de biologie, Ecole Normale Supérieure, CNRS, INSERM, PSL University, 75005, Paris, France

Corresponding Author: mathieu.bahin@biologie.ens.fr

The last ten years have witnessed the rise of a myriad of applications that take advantage of Next-Generation Sequencing (NGS) technologies. In the vast majority of cases, whatever the species, whatever the sequencing technique, the first analysis step of this type of data consists of a quality control of the reads while the second step consists of a mapping of those reads to a reference genome. However, the subsequent steps are often very specific to the type of NGS experiment.

With this work, we aim at introducing a third systematic step after mapping which would be common to any NGS experiment. This step consists in producing a global overview of the distributions of the mapped reads across genomic categories (5'-UTR, CDS, intergenic, stop codon, etc.) and biotypes (protein coding, miRNA, ncRNA, etc.) at nucleotide resolution. Our approach turns out to be very useful for a broad range of NGS applications we are dealing with, as it brings a sort of post-mapping quality control and a first global functional insight. In any case, it adds information to the usual mapped/unmapped read count and other post-mapping statistics.

A few tools providing this type of information have been proposed in the literature for specific NGS applications. For instance, Homer [1] or CEAS [2], dedicated to ChIP-seq data, count detected peaks found in each of a predefined set of categories. However, as those tools cannot conveniently deal with mapped reads, their application to other sequencing techniques is precluded. In fact, to the best of our knowledge, there is no available ready-made tool that proposes such a quantitative overview at a nucleotide precision. Furthermore, using directly the mapped reads allows us to propose a framework working for any species and whatever the sequencing technique.

The tool we propose works in two steps. First, a provided genome annotation file (GTF format) is processed to generate an index. Each nucleotide of the genome is annotated according to a standard priority definition between features. Then the program computes the nucleotide fraction mapped to each predefined feature in one or more BAM files. By default, the program outputs a raw count and a normalized count plots for the categories and another for the biotypes. The normalization is achieved according to the relative importance of a given category or biotype in the genome in order to provide a view in term of enrichment.

We will show results obtained by the proposed tool on various types of NGS experiments such as:

- CLIP-Seq data on Mus Musculus samples to show a rRNA contamination on one sample
- ChIP-Seq data on Caenorhabditis elegans samples to point out a snoNA and rRNA enrichment
- BS-Seq data on Arabidopsis thaliana to discover that some replicates don't show a good reproducibility

References

[1] Heinz S, Benner C, Spann N, Bertolino E, Lin YC, Laslo P, Cheng JX, Murre C, Singh H, Glass CK: Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and B cell identities. Mol. Cell 2010, 38(4):576–589.

[2] Shin H, Liu T, Manrai AK, Liu XS: CEAS: cis-regulatory element annotation system. Bioinformatics 2009, 25(19):2605–2606.

AllMine, a flexible pipeline for allele mining

Thomas Bersez[1,2], Jean-Luc Gallois[1] and Jacques Lagnel[1]
[1] GAFL, INRA PACA, 228 Allée des Chênes, Domaine St Maurice, CS 60094 F-84143 Montfavet Cedex, France
[2] Paris Saclay university, Bat. Discovery RD 128, 91190, Saint-Aubin, France

Corresponding author: thomasbersez@gmail.com

The **wild relatives** of cultivated plants constitute a broad genetic pool of interesting agronomic genetics resources for breeders and agronomists [1,2]. For instance *Solanum lycopersicum* (tomato) register more than a thousand of accessions, wild and field cultivars confounded. This natural reservoir, still largely unexploited, constitute a suitable bank of resistances to pathogens and tolerance to abiotic stress. Furthermore, their closeness with cultivated crops makes gene introgresion feasible.

Ultimately, the bottleneck of this strategy is constituted by our ability to discover variants associated with resistance traits. This process of discovering new alleles of interest is called **Allele Mining** [3]. The complexity of the analysis as well as the numerous steps from raw sequencing reads to *de novo* variants constitute a work-frame for bioinformaticians (whereas biologist are more likely to be involved into the association of *de novo* variants with resistance traits, in the wet-lab). To enable the usage of the complete expense of the modern **NGS** data (*e.i.* read type, pair end vs single end, RNAseq, WGS, RGS ect.), allele mining tools must held an exigence of scalability, modularity and parallelism. Those three points are constitutive of the trinity of pipeline development. Moreover, to allow their usage by the widest public possible, such tools also need to be well documented and implemented in a clear and understandable way.

In such context, we introduce **AllMine**, a flexible pipeline for allele mining. AllMine performs reads prepossessing, mapping, allele mining in defined regions of interest and variant annotation. Also, it can handle various types of inputs such as RNAseq, WGS, paired or single end reads ect. AllMine is designed for **highly parallel computing environments** and so, can fully use computational resources at disposition. AllMine has been implemented using the **Snakemake** workflow manager, in a modular fashion. Furthermore, being deployed in a Singularity container, it does not require dependencies manual installation. Because allele mining AllMine efficiency and accuracy has been verified on both in silico and real data sets collections of variants. Because ambiguous reads can affect variant calling across protein families, AllMine takes account of ambiguous mapping. Outputted variants are tagged as due to uniquely mapped reads, ambiguous reads or both. Discovered variants are presented in an easily browsable spread-sheet.

In the context of the **CASSANDRA consortium project**, AllMine as been tested on *Manihot esculenta* (cassava) sequencing data to look for variants across the **eIF4E** protein family. eIF4E proteins are susceptibility factors to economically important viruses such as Cassava brown streak virus and new eIF4E variants may be associated with genetic resistance.

References

[1] Xiping Yang, Jian Song, Qian You, Dev R. Paudel, Jisen Zhang, and Jianping Wang. Mining sequence variations in representative polyploid sugarcane germplasm accessions. *BMC Genomics*, 18(1), December 2017.

[2] Barbara Hufnagel, Claudia T. Guimaraes, Eric J. Craft, Jon E. Shaff, Robert E. Schaffert, Leon V. Kochian, and Jurandir V. Magalhaes. Exploiting sorghum genetic diversity for enhanced aluminum tolerance: Allele mining based on the AltSB locus. *Scientific Reports*, 8(1), December 2018.

[3] G. Ram Kumar, K. Sakthivel, R.M. Sundaram, C.N. Neeraja, S.M. Balachandran, N. Shobha Rani, B.C. Viraktamath, and M.S. Madhav. Allele mining in crops: Prospects and potentials. *Biotechnology Advances*, 28(4):451–461, July 2010.

An Integrative Deep-Learning Framework for Analyzing Native Spatial Chromatin Dynamics

Hélène Kabbech[1,2], Eduardo Gade Gusmao[1] and Argyris Papantonis[1]
[1] Universitätsmedizin Göttingen, Robert-Koch-Straße 40, 37075, Göttingen, Germany
[2] Université Paris Diderot, 50 rue Alice-Domon et Léonie-Duquet, 75013, Paris, France

Corresponding author: helene.kabbech@gmail.com, eduardogade@gmail.com

The precise spatiotemporal control of gene expression is critical for a cell's correct operation given its identity [1]. Much knowledge was gained by studying gene expression regulatory elements [2]. These elements include, for instance, proteins that bind the DNA to enhance or repress expression of certain genes; and is usually presented in a time-series-like form (2D). Machine learning methods such as hidden Markov models gained an ever-increasing visibility in the computational biology data processing research field [3]. Nevertheless, we have recently come to understand that this control is exerted also via the three-dimensional (3D) organization of the genome (i.e. the chromatin conformation). Thus, the deregulation of gene expression under a state of disease, such as cancer, will most probably involve deregulation of 3D DNA structure [4].

Recently, a number of studies performed integrative analyses using chromatin conformation data and gene regulatory / epigenetic data [5]. However, such studies fail to perform a truly integrative analysis in the sense that they only "overlap" regulatory/epigenetic features at certain chromatin loci. Furthermore, virtually every study so far divides chromatin regions in a discrete manner, i.e. they focus only on compartments, Topologically Associated Domains (TADs), TAD boundaries, contacts called, and a few other discrete structures [6]. Thus, we devised a novel deep-learning-based methodology, which aims to integrate any genomic/regulatory/epigenetic feature (in fact, any biological feature that can be measured along a genomic region) with 3D chromatin conformation data in the form of a contact matrix. The main goal of our methodology is to unravel similar patterns linked to characteristics such as the onset of diseases.

Our Deep Learning framework is composed by a Convolutional Autoencoder (CAE). An Autoencoder (AE) is a type of network that aims to encode an input to a low-dimensional latent space and then decode it back to an output with the same dimensions as the original input. Moreover, AEs are intrinsically self-supervised, i.e. their inputs are also the unmodified targets. CAEs consist of traditional AEs stacked with convolution layers. The idea behind borrowing the convolution operation is to handle: (1) sparsity; (2) biologically- and computationally-derived artifacts and (3) intrinsic heterogeneity [2,3,4]. This is feasible as we are not interested in performing a classification task – but integrating multiple data – thus reducing dramatically the burden of the process of optimizing millions of weights and offset terms in a stacked (deep) architecture, with each layer consisting of a matrix multiplications and offset additions followed by regularization operations. Our ongoing work shows that CAEs are powerful integrative tools.

References

[1] Glenn A. Maston, Sara K. Evans, and Michael R. Green. Transcriptional regulatory elements in the human genome. *Annual Review of Genomics and Human Genetics*, 7(1):29–59, 2006. PMID: 16719718.

[2] R.E. Thurman, E. Rynes, R. Humbert, J. Vierstra, M.T. Maurano, E. Haugen, N.C. Sheffield, A.B. Stergachis, H. Wang, B. Vernot, et al. The accessible chromatin landscape of the human genome. *Nature*, 489(7414):75–82, 2012.

[3] Eduardo G. Gusmao, Manuel Allhoff, Martin Zenke, and Ivan G. Costa. Analysis of computational footprinting methods for DNase sequencing experiments. *Nat Meth*, advance online publication, February 2016.

[4] Alvaro Rada-Iglesias, Frank G Grosveld, and Argyris Papantonis. Forces driving the three-dimensional folding of eukaryotic genomes. *Molecular Systems Biology*, 14(6), 2018.

[5] Hebing Chen, Shuai Jiang, Zhuo Zhang, Hao Li, Yiming Lu, and Xiaochen Bo. Exploring spatially adjacent TFBS-clustered regions with Hi-C data. *Bioinformatics (Oxford, England)*, 33(17):2611–2614, 2017.

[6] Mattia Forcato, Chiara Nicoletti, Koustav Pal, Carmen Maria Livi, Francesco Ferrari, and Silvio Bicciato. Comparison of computational methods for Hi-C data analysis. *Nature Methods*, 2017.

Analyse de longs reads Nanopore avec des k-mers à erreurs

Quentin Bonenfant[1], Laurent Noé[1], Hélène Touzet[1]
CRIStAL, UMR CNRS 9189, Université de Lille, Villeneuve d'Ascq, France

Auteur référent: quentin.bonenfant@univ-lille.fr

Le séquençage de reads longs, avec la technologie Oxford Nanopore par exemple, ouvre de nouvelles perspectives pour la reconstruction des génomes ou l'analyse de transcriptomes. Par rapport aux reads courts, ces reads ont toutefois l'inconvénient majeur de présenter un taux d'erreur élevé (de l'ordre de 10% [1]), ce qui oblige à utiliser des algorithmes spécifiques pour leur traitement.

Les algorithmes développés pour les reads courts font un usage massif d'heuristiques à base de k-mers (mapping avec une BWT, graphes de De Bruijn, ...). Toutefois, l'utilisation de tels k-mers exacts peut entraîner une perte de sensibilité avec des séquences bruitées. Ce problème est d'autant plus patent quand on veut comparer des reads entre eux. Nous proposons l'utilisation de *k-mers avec erreurs* à la place des k-mers exacts pour l'analyse des reads longs. Ces k-mers reposent sur les graines 01^*0 introduites dans [2]. Nous les avons mis en œuvre dans un algorithme d'identification de motifs communs dans un ensemble de reads. La méthode repose sur deux étapes :

- *l'identification* des k-mers composant potentiellement le motif, à l'aide d'une approche par comptage en tenant compte des erreurs,
- *la reconstruction* de la séquence intégrale du motif en utilisant une méthode d'assemblage dans le graphe des k-mers identifiés.

Nous avons appliquée cette approche à la détection des séquences des adaptateurs dans les reads, avec le développement d'une extension de Porechop, dénommée Porechop_ABI (*ab initio*). Cette extension permet d'inférer la séquence de l'adaptateur à partir des reads bruts, pour permettre ensuite le trimming des reads.

Nous avons testé Porechop_ABI sur des données de séquençage d'ADN complémentaire obtenues avec un séquenceur MinION équipé d'une cellule r9.4 en suivant le protocole 1D décrit par Oxford Nanopore. Ce type de reads présentent un taux d'erreur supérieur à 10%. Les résultats montrent un gain net lié à l'utilisation de k-mers avec erreurs. Ces derniers permettent de reconstruire des séquences consensus stables pour 80% des échantillons étudiés, contre 40% seulement avec des k-mers exacts, et ce pour un coût en temps très faible. Porechop_ABI est disponible à https://github.com/qbonenfant/Porechop_ABI à sous licence GPL3.

Acknowledgments : This work was found by ANR (ASTER, ANR-16-CE23-0001).

Références

[1] Miten Jain, John R. Tyson, Matthew Loose, Camilla L.C. Ip, David A. Eccles, Justin O'Grady, Sunir Malla, Richard M. Leggett, Ola Wallerman, Hans J. Jansen, Vadim Zalunin, Ewan Birney, Bonnie L. Brown, Terrance P. Snutch, and Hugh E. Olsen. MinION Analysis and Reference Consortium : Phase 2 data release and analysis of R9.0 chemistry. *F1000Res*, 6, 2017.

[2] Christophe Vroland, Mikaël Salson, Sébastien Bini, and Hélène Touzet. Approximate search of short patterns with high error rates using the 01 0 lossless seeds. *Journal of Discrete Algorithms*, 37 :3–16, 2016.

Analyse du métagénome microbien fonctionnel des sols de parcelles paysannes en zone subsahélienne (Burkina Faso)

Marilyne Aza-Gnandji[1,2], Yves Prin[2], Estelle Tournier[3], Amadou Dieng[1], Ézekiel Baudoin[2], Hervé Sanguin[3], Saliou Fall[1] et Frédéric Mahé[3]

[1] LCM, IRD/ISRA/UCAD, BP 1386 Dakar, Sénégal
[2] CIRAD, UMR LSTM, F-34398 Montpellier, France
[3] CIRAD, UMR BGPI, F-34398 Montpellier, France

Auteur référent: frederic.mahe@cirad.fr

Pour répondre aux défis climatiques et démographiques, une des voies possibles de la transition agricole est l'agroécologie. Celle-ci consiste à mieux prendre en compte et à optimiser les interactions entre plantes cultivées et leur environnement (flore, sol, microbes, climat et sociétés humaines) afin d'améliorer la durabilité des systèmes de culture existants ou à inventer. En Afrique de l'Ouest, l'hétérogénéité des sols cultivés, l'extrême variabilité pluviométrique, l'accès réduit aux intrants, sont également synonymes d'une très large diversité des systèmes de culture. Cette très forte diversité a probablement des effets important sur la diversité et structuration d'une des composantes majeures du fonctionnement du sol : le microbiote. Connaître et caractériser les interactions entre pratiques culturales et fonctionnement biologique du sol est un enjeu majeur.

Notre projet se focalise sur des systèmes de cultures burkinabés associant une forte diversité génétique de sorgho (céréale) et de niébé (légumineuse fixatrice d'azote). Le principal objectif est d'analyser le microbiote des sols de 80 parcelles paysannes réparties sur deux sites au nord de Ouagadougou (Boussouma/Korsimoro et Yilou), le tout via un séquençage total (*Illumina NovaSeq*). Nous présentons ici un premier aperçu de la diversité taxonomique et fonctionnelle de sols agricoles subsahéliens et une comparaison des différents outils bioinformatiques utilisés pour le nettoyage (`vsearch` [1], `cutadapt` [2]), la comparaison (`commet` [3]), l'assignation (`metaxa2` [4]), l'annotation (`prodigal` [5]), l'assemblage (`metaspades` [6]), le *binning* (e.g., `concoct` [7]) ainsi que l'exploration visuelle et statistique des données (`anvi'o` [8]).

Remerciements

Ce travail est financé par la fondation Avril dans le cadre du projet *Oracle*.

Références

[1] Torbjørn Rognes, Tomáš Flouri, Ben Nichols, Christopher Quince, and Frédéric Mahé. VSEARCH : a versatile open source tool for metagenomics. *PeerJ*, 4 :e2584, 2016.

[2] Marcel Martin. Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet.journal*, 17(1) :10–12, 2011.

[3] Nicollas Maillet, Guillaume Collet, Dominique Vannier, Thomas Lavenier, and Pierre Peterlongo. Commet : comparing and combining multiple metagenomic datasets. In *International Conference on Bioinformatics and Biomedicine (BIBM)*. IEEE, 2014.

[4] Johan Bengtsson-Palme, Martin Hartmann, Karl Martin Eriksson, Chandan Pal, Kaisa Thorell, Dan Göran Joakim Larsson, and Rolf Henrik Nilsson. metaxa2 : improved identification and taxonomic classification of small and large subunit rRNA in metagenomic data. *Molecular Ecology Resources*, 15(6) :1403–1414, 2015.

[5] Doug Hyatt, Gwo-Liang Chen, Philip F. LoCascio, Miriam L. Land, Frank W. Larimer, and Loren J. Hauser. Prodigal : prokaryotic gene recognition and translation initiation site identification. *BMC Bioinformatics*, 11(1) :119, 2010.

[6] Sergey Nurk, Dmitry Meleshko, Anton Korobeynikov, and Pavel Pevzner. metaSPAdes : a new versatile de novo metagenomics assembler. *arXiv*, 2016.

[7] Johannes Alneberg, Brynjar Smári Bjarnason, Ino de Bruijn, Melanie Schirmer, Joshua Quick, Umer Z. Ijaz, Leo Lahti, Nicholas J. Loman, Anders F. Andersson, and Christopher Quince. Binning metagenomic contigs by coverage and composition. *Nature Methods*, 11 :1144–1146, 2014.

[8] A. Murat Eren, Özcan C. Esen, Christopher Quince, Joseph H. Vineis, Hilary G. Morrison, Mitchell L. Sogin, and Tom O. Delmont. Anvi'o : an advanced analysis and visualization platform for 'omics data. *PeerJ*, 3 :e1319, oct 2015.

Analysis of multi-omics data: a comparison of correlation and functional integrative approaches on a cancer dataset

Maëlle Daunesse, Vincent Guillemot and Natalia Pietrosemoli
Hub de Bioinformatique et Biostatistique - C3BI , Institut Pasteur, USR 3756 CNRS, Paris, France.

Corresponding author: vincent.guillemot@pasteur.fr and natalia.pietrosemoli@pasteur.fr

As the plethora of techniques to measure biological signal at the molecular level in the same experiment grows, so does the ability to measure simultaneously and on the same sample different types of potentially high dimensional data such as RNA expression, protein abundance, DNA methylation and conformation. Consequently, in the recent years we have witnessed the emergence of novel methods and strategies to jointly analyze the highly heterogeneous types of data resulting from these experiments.

We propose a comparison of two different approaches for the joint analysis of multiple types of omics data: (1) an approach based on the enriched biological functions identified in the different data types, consisting the ranking of the results of an ensemble of nine gene set enrichment tools, and (2) an approach based on the correlation of the different data types based on variations of the Regularized Generalized Canonical Correlation Analysis[1].

Our analysis was performed on a complex dataset of human primary breast cancer [2] from the Cancer Genome Atlas which is often used as an omic analysis case study. From this dataset we selected three blocks of data: i) expression profiles from mRNA arrays and RNAseq (i.e. transcriptomics), ii) microRNA expression (i.e. miRNA transcriptomics) and iii) reverse phase protein arrays measurements (i.e. proteomics) in 348 patients. The samples are classified into five molecular subtypes (groups): four cancer subtypes enriched HER2, basal, luminal A and B and one normal subtype. First, we performed a classical differential analysis using a linear model as implemented in the limma R package [3] followed by a gene set enrichment analysis performed with the EGSEA R package which combines the results of nine independent Gene Set Enrichment Analysis (GSEA) tools [4]. Second, the three blocks of data and the five groups of samples were analyzed with concordance tests and diversity analyses[5]. Third, RGCCA was used to analyze the three blocks of data with (i) a design that reflects an appropriate biological paradigm and (ii) a control design that connects all the blocks. Last, we applied RGCCA to the studentized coefficients generated by the differential analyses to highlight the similarities between the 10 comparisons of two groups.

Our results allow to systematically assess through a wide range of graphical representations and integrated structural information the intrinsic advantages and disadvantages of integrating heterogeneous omic datasets using sophisticated correlation-based methods with respect to using functional analysis analysis methods based on the shared biological context of the genes. Such context might include shared molecular pathways, biological processes and cellular compartments, among others.

References

[1] Arthur Tenenhaus and Michel Tenenhaus. Regularized Generalized Canonical Correlation Analysis. *Psychometrika*, 76(2):257–284, mar 2011.

[2] Daniel C. Koboldt, Robert S. Fulton, Michael D. McLellan, Heather Schmidt, and et al. Kalicki-Veizer. Comprehensive molecular portraits of human breast tumours. *Nature*, 490(7418):61–70, sep 2012.

[3] Matthew E. Ritchie, Belinda Phipson, Di Wu, Yifang Hu, Charity W. Law, Wei Shi, and Gordon K. Smyth. limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Research*, 43(7):e47–e47, apr 2015.

[4] Monther Alhamdoosh, Milica Ng, Nicholas J. Wilson, Julie M. Sheridan, Huy Huynh, Michael J. Wilson, and Matthew E. Ritchie. Combining multiple tools outperforms individual methods in gene set enrichment analyses. *Bioinformatics*, page btw623, September 2016.

[5] Vamsi K. Mootha, Jakob Bunkenborg, Jesper V. Olsen, Majbrit Hjerrild, and Wisniewski et al. Integrated Analysis of Protein Composition, Tissue Diversity, and Gene Regulation in Mouse Mitochondria. *Cell*, 115(5):629–640, nov 2003.

Analysis workflow for low frequency variant detection

Xavier MIALHE[1], Stéphanie RIALLE[1], Marine PRATLONG[1] and Emeric DUBOIS[1]
Plateforme MGX - Montpellier GenomiX, c/o Institut de Génomique Fonctionnelle, 141, rue de la cardonille, 34094 MONTPELLIER Cedex 05, FRANCE

Corresponding author: stephanie.rialle@mgx.cnrs.fr

1 Abstract

Next Generation Sequencing data analysis is a constantly evolving field. The detection of genetic variants from sequencing data is a complex problem.

If the search for variants is easier for germinal variants, one important current research field, cancer, is more towards the search for somatic variations that have a low appearance frequency compared to germinal variants.

Highlighting somatic variants is complicated due to the high rate of false-positives detection. PCR artifacts, sequencing errors or sample degradation (Formalin-Fixed Paraffin-Embedded tissues for exemple) can introduce a bias in this detection and statistical tests or error correction techniques do not always enable to remove these false-positive results.

After reviewing the state of the art of existing somatic variant callers, the developement of a benchmark of these tools (VarDict [4] and Octopus [1] for exemple) on simulated data with BAMSurgeon [2] and the implementation of the best tool, we introduce you a Snakemake [3] workflow for low frequency variant detection running with paired or single-end data from Illumina technology.

The workflow is used on the Montpellier GenomiX facility to provide a new analysis service.

References

[1] Daniel P. Cooke, David C. Wedge, and Gerton Lunter. "A unified haplotype-based method for accurate and comprehensive variant calling". In: *bioRxiv* (Oct. 29, 2018), p. 456103. DOI: 10.1101/456103. URL: https://www.biorxiv.org/content/10.1101/456103v1.

[2] Adam D. Ewing et al. "Combining tumor genome simulation with crowdsourcing to benchmark somatic single-nucleotide-variant detection". In: *Nature Methods* 12.7 (July 2015), pp. 623–630. ISSN: 1548-7105. DOI: 10.1038/nmeth.3407. URL: https://www.nature.com/articles/nmeth.3407.

[3] Johannes Köster and Sven Rahmann. "Snakemake a scalable bioinformatics workflow engine". In: *Bioinformatics* 28.19 (Oct. 1, 2012), pp. 2520–2522. ISSN: 1367-4803. DOI: 10.1093/bioinformatics/bts480. URL: https://academic.oup.com/bioinformatics/article/28/19/2520/290322.

[4] Zhongwu Lai et al. "VarDict: a novel and versatile variant caller for next-generation sequencing in cancer research". In: *Nucleic Acids Research* 44.11 (June 20, 2016), e108–e108. ISSN: 0305-1048. DOI: 10.1093/nar/gkw227. URL: https://academic.oup.com/nar/article/44/11/e108/2468301.

Apollo method: statistical inference to reveal hidden data in chromosomal contact maps

AXEL COURNAC[1]

[1] Institut Pasteur, Unité Régulation Spatiale des Génomes, 28 rue du Docteur Roux, 75015,Paris, France

Corresponding Author: acournac@pasteur.fr

Abstract *The 3D structure of chromosomes may impact or be impacted by major biological functions such as replication, segregation or transcription. To observe and study spatial organization of chromosomes, so-called contact techniques (3C, Hi-C) are developed in parallel with microscopy. They are based on the capture and quantification of physical contact between different loci within a genome and bring a new type of information to an unprecedented spatial resolution. These techniques can generate millions pairs of short sequences (~ 50 nucleotides), a certain proportion of which cannot be located directly due to their repetition in the sequence of the reference genome (several alignments are possible). To overcome this limitation, we propose the Apollo method, which uses statistical inference to predict the contacts of the repeated sequences and thus reveal the hidden side of chromosomes. Unpublished results will be presented with applications on micro-organisms contact maps like Escherichia coli, Vibrio Cholerae bacteria or yeast Saccaromyces cerevisiae.*

Keywords Chromosome organisation- repeated sequences – statistical inference – contact data – HiC, 3C – Microbiology -

1 Biological context

Links between structure and functions are very common in Biology: the precise protein tertiary structure dictates its activity. The precise folding of RNA molecules can give them regulatory properties. Another connection between 3D structure also occurs at the chromosome level. Indeed, it is now becoming increasingly clear that the precise architecture of chromosomes underlies major biological functions such as replication, segregation or transcriptional regulation.

To observe and study this specific organization, contact techniques have been proposed and are currently in full expansion in parallel of microscopy. These technologies are based on the capture and quantification of physical contact between different loci within a genome [1] and bring a new type of information to an unprecedented spatial resolution. These techniques have been used on a very wide variety of organisms notably bacteria, yeast and metazoans and have revealed several levels of organization.

Human genome is partitioned into two compartments; active and inactive [2] that correlate with either open and closed chromatin. Recent studies propose that phase separation could be a mechanism that mediates this genome organisation [3].

Chromosomes are organized into domains that preferentially self-interact called Topologically Associated Domains (TADs) [4]. The molecular mechanisms behind the formation of these TADs may imply active process called loop extrusion involving condensins or cohesin proteins that progressively loop extrude DNA with energy consumption [5]. Interesting, these mechanisms may be universal and also apply to bacteria, yeast and human chromosomes [6].

These techniques also enable the fine observation of genomic loops (with size of ~100 kb) linking notably enhancers and promoters by specific Transcription Factors (TF) showing with unprecedented resolution the links between chromosome architecture and gene regulation during cell differenciation [7] or autoimmune disease [8].

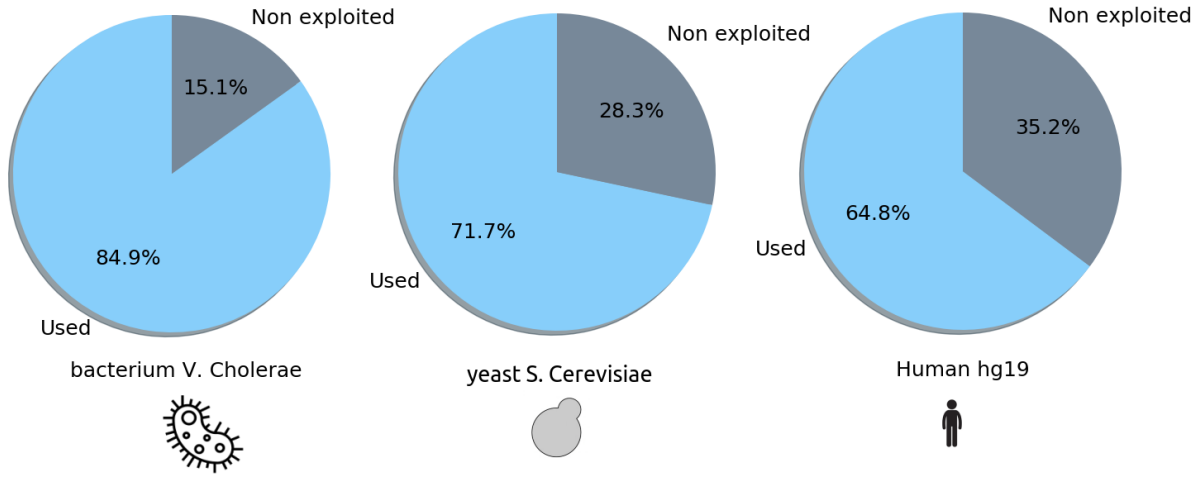
2 New hypothesis and motivation

Beside specific proteins, we can also propose that mechanisms involving repeated elements may also contribute to the structuring of chromosomes. Interestingly, there is more and more evidence showing that homologous (or similar) sequences could make preferential contacts. Single molecule experiments show direct homologous, DNA/DNA pairing [9], bioinformatics analyses suggest colocalisation of repeated elements in metazoan genomes [10], theoretical physics models propose subtle mechanisms of homologous pairing through the formation of short quadruplexe [11]. Recently, experimental study on the fungus Neurospora crassa showed that repeat-induced point mutation (RIP) involves direct interactions between homologous double-stranded DNA segments [12] and that recognition depends on the positions of mutations suggesting that the periodicity and physics of DNA is important in the repeat recognition process.

These diverse results from different scientific communities are compatible with the hypothesis that repeated sequences can play a direct role in genome architecture by notably making specific contacts which represents a very exiting and timely model to test [13]. We can also imagine that certain repeated make other contact patterns like domain boundaries. The proposed method aims to understand the general impact of repeated sequences on the spatial organization of genomes.

3 Current limitations of the computational methods for Hi-C data

One of the limitation of the current Hi-C technologies is a common limitation encountered in any pipelines using Next Generation Sequencing (NGS) data. The read size (for example:~ 50 bp) can generate ambiguous mapped positions during the alignment procedure i.e one sequence can be located in several possible positions in the reference genome. This category of sequences are currently discarded with standard pipelines (for a recent review see [14]). The percentage of such sequences can represent 35% of a common Hi-C human library (Fig. 1.) preventing the observation and analysis of contacts involving repeated sequences. There is thus a need for the development of new algorithms to use these sequences and extract biological information from them. At the time writing this JOBIM proposal, if we consider only the human Hi-C datasets available on public server like Sequence Read Archive of NCBI, we have ~ 2000 libraries (done in various biological conditions and cell types) corresponding to ~ 200 billion of pairs of sequence. If we consider that about 35% of these data are eliminated using the standard computational methods, it corresponds to the order of hundreds of terabits of genomic data that are not currently exploited.



In

Fig 1. *Percentage of pairs of reads (i.e with Mapping Quality above 30) that are kept for subsequent processing of contact data (Hi-C, 3Cseq) with current pipelines.*

Interestingly, the hidden information contained in those filtered reads contain contact signal about singular genomic objects composed of repeated DNA that belong to “the far side of the chromosome”.

4 Implementation

To illustrate the feasibility of our approach, we propose in this section several features that could be used for the implementation of the Apollo method. Besides using machine learning approaches on simulated data, an interesting possibility is to carry out the training step on the visible part on the genome. This implies that the *multi-mappable part of the genome will behave the same way as the mappable part on certain features.* In a first approximation, this hypothesis can be acceptable for several biological and physical signals, we give two examples. The first one is the general coverage along the genome that will depends for bacteria on the replication timing. We give an illustration of such signal on Fig 2.A representing the replication timing of Escherichia coli genome. *Hi-C* coverage is stronger near the ori of replication and weaker near the terminus of replication noted ter. The multi-mappable reads will have in consequence more chance to come from Ori than ter and the replication timing can be used as a first probability law of read prediction. The second feature that can be used is the probability of contact in function of the genomic distance represented in Fig2.B.

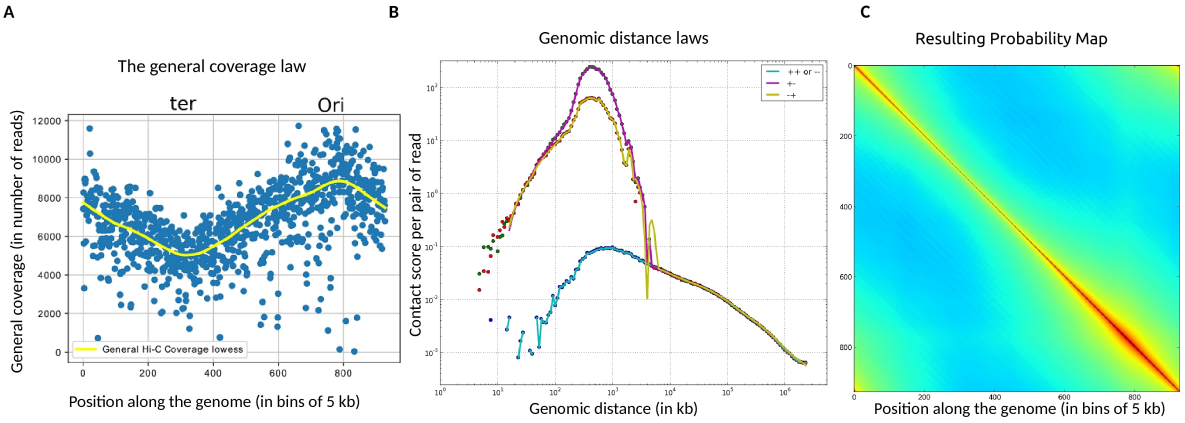


Fig. 2. *Examples of biological and physical signals present in the visible part of the data that can be used to infer the positions of currently multi-mappable positions in chromosomal contact maps. (A) The general coverage signal corresponds to the replication timing in Escherichia coli genome. Data were binned at 5 kb and smoothed with a lowess procedure. (B) Probability of contact depending on the genomic distance for different configurations of pairs of reads taking into account the direction of the read on the reference genome [noted + or -]. (C) Combining replication timing signal and probability of contact in function of the genomic distance results in a Probability Map to detect each pair of loci in the chromosomal contact map.*

This function is exponentially decreasing and is related to the polymer nature of chromosomes. This law can be carefully computed for each configuration of read pairs with their direction of mapping on the reference genome. Indeed, these different configurations correspond to particular events occurring in a *Hi-C* or *3C-seq* library like undigested co-linerar fragments, re-circularised fragments etc that we previously described [18]. The knowing of distribution of these different event is crucial for the correct assignment of multi-mappable reads. The latter two laws can be combined to build in a first approximation of a Probability Map (Fig4.C) that gives the expected detection of each pair of reads in the contact map of E.Coli. This map can then be used to reassign all potential positions coming from ambiguous reads present in the Hi-C library. As proof of concept of our method, we give two biological analyses using this approach in the next section, some of them with preliminary and encouraging results. Integration of other biological data can guide as well the correct prediction of contacts. For example, we recently detected a positive correlation between transcription level and short range contacts in bacteria [19]. We also detected correlation between cumulative contact signal and mobilities measurements done in microscopy [19].

5 Biological applications

5.1 Application 1: ribosomal operons network of contacts in Escherichia coli genome

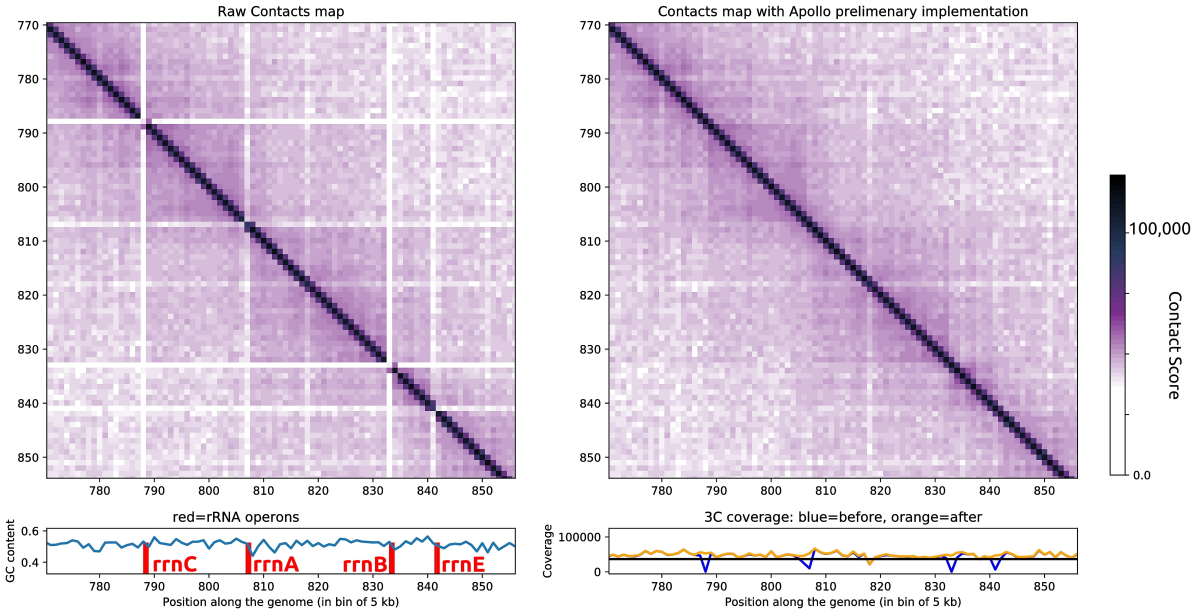


Fig. 3. *Zoom of the contact map from the Escherichia coli genome involving 4 rDNA operons [A]before and (B) after reconstruction with preliminary Apollo implementation. (C) The rDNA are represented by red boxes below the contact map with GC content signal. (D) 3C coverage before [blue]and after [orange]the Apollo reconstruction. After reconstruction, the 4 rDNA recover a normal coverage.*

Ribosomal RNAs in E.coli are transcribed from seven operons, which are highly conserved in their organization and sequence. They are positioned at 7 different loci on the chromosome, located near the origin of replication. They are composed of 3 genes (coding for subunits of ribosomal RNA). The intergenic regions are not exactly identical and allow some mappability but they are of small size. They are the most transcribed genes in the genome of E.Coli and expressed several orders of magnitude than the rest of the transcriptome. They are constantly expressed and the translation of all proteins depends on them. These features are shared in many different bacteria like *Bacillus subtilis, Vibrio Cholerae* and *Pseudomonas aeruginosa*. Interesting, is has been speculated that these operons form a singular spatial structure in the bacterial cell, a kind of nucleolus where the genes are transcribed in the same place inside the cell [15]. Recent microscopy measurements of inter-focal distances for several pairs of rDNA from the laboratory of R. Gourse found that all but rrnC are in close proximity inside the cell [16].

Fig.3.B shows a first reconstruction of contact map involving 4 rRNA at the end of the E.Coli genome. The recreated contact signal does not show enrichment pattern between different rDNA for the moment. This first reconstruction may raise a contradiction between contact data and microscopy measurements. A possible explanation is that cross-linking procedure currently used in the majority of Hi-C protocols (based on formaldehyde agent) does not allow to detect these interactions, other crosslinkers (like DMA or EGS with longer arm size) may be necessary. The data presented in Fig.3 correspond to bacteria culture in minimal medium with a rather slow growth. It would be very interesting to apply the reconstruction in different culture conditions notably rich medium (like LB), other temperature (these data are available for E.Coli and other bacteria). Finally, we cannot also exclude that this first implementation of Apollo method is not able for the moment to predict enrichment patterns. Additional ingredients in the procedure like local 2D interpolation (using mappable flanking regions) may be necessary and implemented. Tests on simulated chromosome will help confirming each hypothesis.

5.2 Application 2: Superintegron present in Vibrio Cholerae chromosome 2

Superintegrons are very singular and interesting genetic objects present notably on the chromosome 2 of *Vibrio Cholerae* (and on other bacteria). It is a region composed of a cluster of repeated sequences with the presence of specific integrase and recombination sites [17]. It has the unique property to incorporate exogenous open reading frames and convert them into functional genes by ensuring their correct expression [17]. They can be compared to *assembly platforms*.

After Apollo preliminary reconstruction, the Superintegron recovers a normal coverage (even a bit more covered compared to the chromosome average, Fig4.D) however it still does not seem to contact other regions of the chromosome. It seems that it behaves in a very isolated manner and does not interact with the rest of the genome. Several interpretations are possible: it can be due to a current limit of the Hi-C protocol; GC content of this region is very different from the rest of the genome (Fig. 4.C) and the restriction sites density could distort the contact signal. A biological explanation would be that proteins that bind the Superintegron and/or local chromatin organisation of this object are very different of the rest of the genome so physical and stable contacts are depleted between them. A confrontation with microscopy data will be relevant and bring information in favour of one or other hypothesis. Another striking event of the reconstructed contact map is the presence of a stripe at a specific locus (represented with a black arrow on Fig4.B).

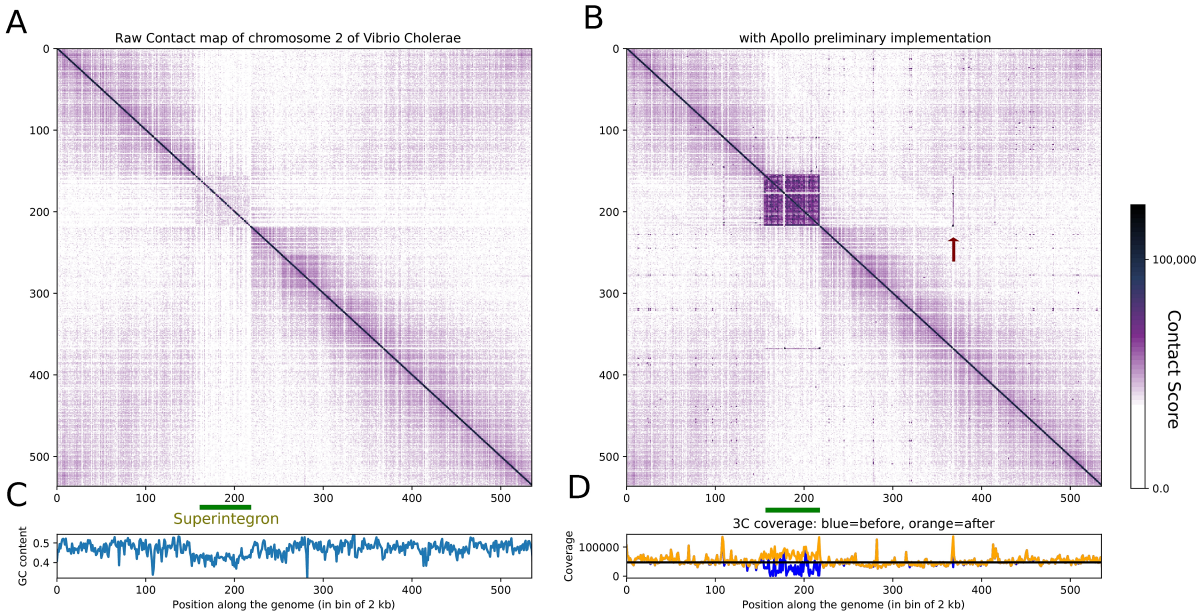


Fig. 4. *Contact map from the Vibrio Cholerae chromosome 2 involving Superintegron object (A) before and (B) after reconstruction with preliminary Apollo implementation. (C) The Superintegron is represented with orange rectangle below the contact map with GC content signal. (D) 3C coverage before (blue) and after (orange) the Apollo reconstruction. After reconstruction, the Superintegron recovers a normal coverage.*

This signal may not be compatible with real physical contact but may be due to sequence change with the reference genome. Here the sequence localised at the arrow on the contact map may have been absorbed by the Superintegron and should be positioned inside it. This second example also underlines that the Apollo approach by carefully and systematically analysing all the contact signals of exotic sequences will certainly bring new information about genome plasticity in micro-organism chromosomes.

6 Conclusion

The predictions of the Apollo algorithm should bring new hypothesis/information concerning the potential roles of repeated elements on genome 3D structure notably spatial organisation of various singular objects present in prokaryotes genomes but also the role of certain transposable elements in metazoan genomes (notably from the Alu family [10]) as potential alternative enhancers in differentiation/ cancerisation / ageing of cells. We think that the approach may also bring unexpected information about genome sequence plasticity connected to repeated elements thought the lens of contact data. The method we propose will also be of great interest by improving the homogeneity of the data. Working with more complete data will also improve the quality of the already visible part of the genome giving more evenly distributed signal at the genome scale. It will improve the normalization procedure and automatic detection of specific contact patterns like domains or loops. It can also challenge our current Hi-C protocol by highlighting contradiction between different techniques.

Acknowledgements

We thank all the members of Romain Koszul laboratory for stimulating and fruitful discussions.

References

[1] Dekker J, Rippe K, Dekker M, Kleckner N. Capturing chromosome conformation. Science. 2002 Feb 15;295[5558]:1306–11.

[2] Lieberman-Aiden E, van Berkum NL, Williams L, Imakaev M, Ragoczy T, et al. Comprehensive mapping of long-range interactions reveals folding principles of the human genome. Science. 2009 Oct 9;326[5950]:289–93.

[3] Strom AR, Emelyanov AV, Mir M, Fyodorov DV, Darzacq X, Karpen GH. Phase separation drives heterochromatin domain formation. Nature. 2017 Jul;547[7662]:241–245.

[4] Dixon JR, Selvaraj S, Yue F, Kim A, Li Y, Shen Y, et al. Topological domains in mammalian genomes identified by analysis of chromatin interactions. Nature [Internet]. 2012 May;485[7398]:376–380.

[5] Fudenberg G, Imakaev M, Lu C, Goloborodko A, Abdennur N, Mirny LA. Formation of chromosomal domains by loop extrusion. Cell reports. 2016;15[9]:2038–2049.

[6] Marbouty M, Le Gall A, Cattoni DI, Cournac A, Koh A, Fiche J-B, et al. Condensin- and Replication-Mediated Bacterial Chromosome Folding and Origin Condensation Revealed by Hi-C and Super-resolution Imaging. Molecular Cell [Internet]. 2015 Aug;59[4]:588–602.

[7] Bonev B, Mendelson Cohen N, Szabo Q, Fritsch L, Papadopoulos GL, Lubling Y, et al. Multiscale 3D Genome Rewiring during Mouse Neural Development. Cell. 2017 Oct 19;171[3]:557-572.e24.

[8] Burren OS, Rubio García A, Javierre B-M, Rainbow DB, Cairns J, Cooper NJ, et al. Chromosome contacts in activated T cells identify autoimmune disease candidate genes. Genome Biol. 2017 04;18[1]:165.

[9] Danilowicz C, Lee CH, Kim K, Hatch K, Coljee VW, Kleckner N, et al. Single molecule detection of direct, homologous, DNA/DNA pairing. Proceedings of the National Academy of Sciences of the United States of America. 2009 Nov;106[47]:19824–19829.

[10] Cournac A, Koszul R, Mozziconacci J. The 3D folding of metazoan genomes correlates with the association of similar repetitive elements. Nucleic acids research. 2016;44[1]:245–255.

[11] Mazur AK. Homologous Pairing between Long DNA Double Helices. Physical review letters. 2016 Apr;116[15]:158101.

[12] Gladyshev E, Kleckner N. Recombination-Independent Recognition of DNA Homology for Repeat-Induced Point Mutation [RIP]Is Modulated by the Underlying Nucleotide Sequence. PLoS genetics. 2016 May;12[5]:e1006015.

[13] Cournac A. Aspects temporel et spatial dans des systèmes de régulation génétique [PhD Thesis]. Université Paris-Diderot - Paris VII; 2009.

[14] Forcato M, Nicoletti C, Pal K, Livi CM, Ferrari F, Bicciato S. Comparison of computational methods for Hi-C data analysis. Nature methods. 2017;14[7]:679.

[15] Lewis PJ, Thaker SD, Errington J. Compartmentalization of transcription and translation in Bacillus subtilis. The EMBO Journal [Internet]. 2000 Feb 15 [cited 2019 Mar 25];19[4]:710–8.

[16] Gaal T, Bratton BP, Sanchez-Vazquez P, Sliwicki A, Sliwicki K, Vegel A, et al. Colocalization of distant chromosomal loci in space in E. coli: a bacterial nucleolus. Genes Dev. 2016 Oct 15;30[20]:2272–85.

[17] Mazel D. Integrons: agents of bacterial evolution. Nat Rev Microbiol. 2006 Aug;4[8]:608–20.

[18] Cournac A, Marbouty M, Mozziconacci J, Koszul R. Generation and analysis of chromosomal contact maps of yeast species. Yeast Functional Genomics: Methods and Protocols. 2016;227–245.

[19] Lioy VS, Cournac A, Marbouty M, Duigou S, Mozziconacci J, Espéli O, et al. Multiscale Structuring of the E. coli Chromosome by Nucleoid-Associated and Condensin Proteins. Cell. 2018 Feb 8;172[4]:771-783.e18.

Assessment of inflammatory and immune pathways in Rheumatoid Arthritis patients using BIOPRED kit

Sami AIT ABBI NAZI [1], Eric SCHORDAN [1], [1] and Huseyin FIRAT [1]

[1] Firalis, 35, rue du Fort, 68330 Huningue, France

Corresponding Author: sami.nazi@firalis.com

1. Background

Rheumatoid Arthritis (RA) is a chronic, progressive, inflammatory autoimmune disease associated with articular, extra-articular and systemic effects leading to joint destruction. T cells, B cells and the orchestrated interaction of pro-inflammatory cytokines play key roles in the pathophysiology of RA. Better comprehension of interaction between cytokines and their signaling pathways are key for the development of new strategies with small molecules or biologicals. Today, new technologies allow the specific investigation of inflammatory pathways on mRNA level without extraction step directly from the blood. The BIOPRED panel, based on HTG EdgeSeq platform is a targeting sequencing panel with focus on specific biological pathways including 2155 mRNA from inflammatory and immune pathways. Disease activity specific gene enrichment studies in Rheumatoid Arthritis (RA) & other autoimmune-inflammatory disorders would help pharmaceutical industry tailor pathway specific therapies and would help clinicians choose optimal & personalized therapy for their patients. Therefore, we have identified active biological pathways in RA patients associated with different disease activity status.

2. Objectives

By using Firalis' BIOPRED panel, an innovative targeted gene sequencing panel of 2155 mRNA targets associated with immune-inflammatory pathways, our objective is to identify active biological pathways in function to different disease activity status of RA & Healthy volunteer (HV) subjects.

3. Methods

Paxgene samples of active RA patients with DAS28>3.2 (n= 178) and HV (n= 25) are directly profiled without RNA extraction with BIOPRED panel on HTG EdgeSeq platform, a combination of a nuclease protection assay & next generation sequencing (NGS). Subjects are categorized into three groups; High disease activity (HDA) group with DAS28 >5.1, Moderate disease activity (MDA) group with DAS28 between 3.2 and 5.1, and Healthy volunteer (HEV) group.

4. Results:

After transformation and normalization of the gene expression data, 22 mRNA genes are found to be significantly upregulated in RA (p-value < 0.005, fold change > 2) as compared to HV group. After one-way ANOVA analysis on three groups as stated above, 351 mRNA targets are significantly regulated (p-value <0.05). Pathway analysis based on protein-protein interaction from Biogrid, String and Intact database was used and score were generated based on fold change for 22 pathways to assess modification in HDA and MDA groups versus HV group. Various pathways including Jak/STAT pathway were shown to be significantly upregulated.

5. Conclusions

Our results identify a list of mRNAs relevant to RA pathology and pathways and have the potential to be candidate biomarkers/therapeutic targets. Moreover, BIOPRED panel accurately measures 2155 mRNA

from inflammatory and immune pathways and can be further used to study pathway analysis in autoimmune-inflammatory disorders such as RA.

BamCramConverter: Utility for Easy Alignment/Map Data Storage

David Baux[1], Michel Koenig[1] and Anne-Françoise Roux[1]
Laboratoire de génétique moléculaire des maladies rares, EA7402, Université de Montpellier, Laboratoire de génétique moléculaire, CHU de Montpellier, Université de Montpellier, Montpellier, France

Corresponding author: david.baux@inserm.fr

Current sequencing technologies produce large amount of data, especially alignments of reads which can be challenging to store efficiently. Several formats have been designed to save space and store sequences derived from or more or less compatible with the Sequencing Alignment Map (SAM) format which describes sequences, mapping and quality values [1]. The most widely used is the Binary Alignment Map (BAM) representation of the data [1], but the Compressed and Reference-oriented Alignment Map (CRAM) specification [2] is becoming popular as it is more efficiently compressed than the BAM format. In addition it is directly handled by many popular tools such as samtools (via htslib), and, via htsjdk by the Genome Analysis ToolKit (GATK) or the Integrative Genomics Viewer (IGV) software for visualization. Recently, Crumble [3], a new compression method of quality values, has been described, and is compatible with both BAM and CRAM files.
Applying a clear strategy for file conversion/compression to a complete storage unit involves several steps which can quickly become fastidious if not automated:

- Identification of candidate files
- File conversion
- File indexing
- Consistency checking
- Re-compression using Crumble
- Re-indexing
- Removal of original files

We present the BamCramConverter utility, a bash script integrating UNIX find, samtools, bam2cram-check and Crumble that is flexible enough to automate all these tasks for the end-user. BamCramConverter uses UNIX find on a given directory to select the files that match a given format (BAM/CRAM), a given size and last modification time (provided by the user as arguments). It is able to run sequentially on these files samtools to convert and index BAM/CRAM files in both directions.
On demand, or automatically if the user asks for original files deletion, it can also use a slightly modified version of bam2cram-check which verifies the consistency between the original and the converted files. Optionally, it also manages Crumble to further reduce file sizes. Using Crumble, we observed on our data a 40% average of file size reduction for big BAM files (e.g. from 15Go to 8.5Go) and up to 80% for small gene panels CRAM files (e.g. from 265Mo to 51Mo).

In addition, BamCramConverter can manage multithreading modes of samtools and Crumble, and can be used in SLURM environment by generating srun commands. Finally, it includes a dry-run mode for safety. BamCramConverter offers an efficient solution to automate and reduce space storage to any institution generating Illumina sequencing data.

BamCramConverter is available on GitHub under a GNU GPL v3.0 license.

References

[1] Heng Li, Bob Handsaker, Alec Wysoker, Tim Fennell, Jue Ruan, Nils Homer, Gabor Marth, Goncalo Abecasis, and Richard Durbin. The Sequence Alignment/Map format and SAMtools. *Bioinformatics*, 25(16):2078–2079, August 2009.

[2] Markus Hsi-Yang Fritz, Rasko Leinonen, Guy Cochrane, and Ewan Birney. Efficient storage of high throughput DNA sequencing data using reference-based compression. *Genome Research*, 21(5):734–740, May 2011.

[3] James K Bonfield, Shane A McCarthy, and Richard Durbin. Crumble: reference free lossy compression of sequence quality values. *Bioinformatics*, 35(2):337–339, January 2019.

Benchmarking Hi-C scaffolders

Nadège Guiglielmoni[1], Romain Koszul[2] and Jean-François Flot[1,3]
[1] Service Evolution Biologique et Ecologie, Université libre de Bruxelles, 1050 Brussels, Belgium
[2] Equipe Régulation Spatiale des Génomes, Institut Pasteur, 75015 Paris, France
[3] Interuniversity Institute of Bioinformatics in Brussels - (IB)[2], 1050 Brussels, Belgium

Corresponding author: nadege.guiglielmoni@ulb.be

1 Introduction

New sequencing technologies have multiplied over the past decades and fully assembled genomes are now achieved. At present, sequencing strategies typically involve a combination of techniques such as short-reads, long-reads and further experiments to fill the gaps and obtain reliable genomes. Chromosome conformation capture (3C) [1] was originally developed to study 3D contacts in DNA, and its latest improvement, Hi-C [2], is able to document the conformation of a genome in its entirety. Due to the physical properties of DNA, the 3D structure of chromatin is dependant on its 1D sequence: contacts in 3D occur more frequently between loci that are close along the DNA sequence [2]. Various tools have consequently been developed to make use of Hi-C data to generate chromosome-length scaffolds [3]. For this purpose, the different scaffolders available use strikingly different approaches: 3D-DNA [4] and SALSA [5] are based on graph theory and uses Hi-C links to join contigs, whereas GRAAL [6], further developed as instaGRAAL, reassembles genomes into the structure most likely to explain the observed interaction frequencies. Although these tools have all proved their efficiency to produce less fragmented assemblies, it is unclear how they behave depending on the type of genome, and the errors they can make. We are therefore benchmarking these Hi-C scaffolders on genomes for which good reference assemblies and Hi-C data are already available. Our benchmark will help users adapt their strategy to the characteristics of the genome they want to assemble.

2 Material and Methods

We are testing Hi-C scaffolders on simulated fragmented assemblies using published Hi-C data. We are focussing on two model species: *Caenorhabditis elegans*, whose genome has already been fully assembled; *Drosophila melanogaster*, which displays a more complex genomic structure.

Acknowledgements

This project is funded by the European Union's Horizon 2020 research and innovation program under the Marie Skłodowska-Curie grant agreement 764840.

References

[1] Job Dekker, Karsten Rippe, Martijn Dekker, and Nancy Kleckner. Capturing chromosome conformation. 295(5558):1306–1311, 2002.

[2] Erez Lieberman-Aiden, Nynke L van Berkum, Louise Williams, Maxim Imakaev, Tobias Ragoczy, Agnes Telling, Ido Amit, Bryan R Lajoie, Peter J Sabo, Michael O Dorschner, Richard Sandstrom, Bradley Bernstein, M A Bender, Mark Groudine, Andreas Gnirke, John Stamatoyannopoulos, Leonid A Mirny, Eric S Lander, and Job Dekker. Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *Science*, 326(5950):289–93, 2009.

[3] Jean-François Flot, Hervé Marie-Nelly, and Romain Koszul. Contact genomics: scaffolding and phasing (meta)genomes using chromosome 3D physical signatures. *FEBS Letters*, 589(20):2966–2974, 2015.

[4] Olga Dudchenko, Sanjit S. Batra, Arina D. Omer, Sarah K. Nyquist, Marie Hoeger, Neva C. Durand, Muhammad S. Shamim, Ido Machol, Eric S. Lander, Aviva Presser Aiden, and Erez Lieberman Aiden. De novo assembly of the *Aedes aegypti* genome using Hi-C yields chromosome-length scaffolds. *Science*, 356(6333):92–95, 2017.

[5] Jay Ghurye, Mihai Pop, Sergey Koren, Derek Bickhart, and Chen Shan Chin. Scaffolding of long read assemblies using long range contact information. *BMC Genomics*, 18(1):1–11, 2017.

[6] Hervé Marie-Nelly, Martial Marbouty, Axel Cournac, Jean François Flot, Gianni Liti, Dante Poggi Parodi, Sylvie Syan, Nancy Guillén, Antoine Margeot, Christophe Zimmer, and Romain Koszul. High-quality genome (re)assembly using chromosomal contact data. *Nature Communications*, 5:1–10, 2014.

Bioanalysis activities on the ABiMS (Analysis and Bioinformatic for Marine Science) platform

Enora GESLAIN[1], Julien ROBERT[1], Jacky AME[1], Erwan CORRE[1] and our dear collaborators.

[1] CNRS - Sorbonne Université - Plateforme ABIMS - Station Biologique de Roscoff, Place Georges Teissier, 29680, Roscoff, France

Corresponding Author: erwan.corre@sb-roscoff.fr

The mission of the ABiMS platform is to assist researchers of the marine community and, more broadly, of the life sciences, in the bioinformatic analysis of their data as well as in the development of software and databases. It is one of the national platforms of the French Institute of Bioinformatics (IFB). It is also associated to the EMBRC (European Marine Biology Ressource Center) infrastructure, is part of the IBiSA network via the regional BioGenouest project and is ISO 9001: 2015 certified. Through its numerous interactions with research units, ABiMS is involved in several projects, with national and European impacts involving bioanalysis activities, software, and e-Infrastructures development. Through 3 examples of collaborative projects conducted by Bachelor and Master students we wish to illustrate the bioanalysis activity conducted by the ABiMS platform in the field of marine data.

Algavor project (collaboration with F. Thomas – UMR8227): The recycling of macroalgal biomass influences the functioning of coastal ecosystems. It relies heavily on pioneer bacteria capable of attacking intact algal tissues and releasing degradation products into the water column. As part of this project, we are exploring the presence of some of these pioneer bacteria of the genus *Zobellia* in marine, coastal or alga-associated metagenomes. We use available genomes of pure *Zobellia* strains to recruit reads and evaluate the distribution, abundance, and activity of *Zobellia* spp. in marine environments. Further, we will attempt to build metagenome-assembled genomes (MAG) from recruited reads to gain insights into their biodiversity and catabolic functions.

Algal holobiont project (collaboration with E. Karimi and S. Dittami – UMR8227): Brown algae form key marine ecosystems and live in tight relationship with bacteria essential to their growth and development. Environmental changes can cause imbalances in these systems, leading to shifts in bacterial communities and sometimes giving rise to pathogenic interactions. Hence, considering both the algae and their microbiome (*i.e.* the holobiont) is of essence to understand their response to environmental challenges. In this context, 60 strains of microorganisms were isolated from algal tissue and algal growth media. Their genomes were sequenced, assembled, and different automatic annotation pipeline were evaluated. Finally, their metabolic networks have been reconstructed to estimate their degree of complementarity and thus the potential for beneficial interactions with the metabolic network of their algal host.

Honeycomb worm metatranscriptome analysis (collaboration with F. Nunes - IFREMER): The reef building honeycomb worm, *Sabellaria alveolata,* is an ecosystem engineer that supports high biodiversity, yet there are currently no genomic resources available for this species. In this study, we present the assembly of an annotated *de novo* meta-transcriptome for *S. alveolata*, based on RNASeq of 75 whole individuals from 5 populations across the species' range exposed to five temperature treatments. We focus our study on a detailed annotation of host and non-host transcripts, as well as on the biogeographical patterns of the eukaryotic fauna associated with the worms.

Bioconvert, a common bioinformatics format converter library: status and perspectives.

Sulyvan Dollin[1], Bertrand Néron[1]*, Thomas Cokelaer[1,2]*

[1] Institut Pasteur – Bioinformatics and Biostatistics Hub – C3BI, USR 3756 IP CNRS – Paris, France

[2] Institut Pasteur – Platform Biomics – 25-28 Rue du docteur Roux, Paris, France.

Corresponding Author: thomas.cokelaer@pasteur.fr

Abstract

Life sciences involve the knowledge and use of many different data formats. Their diversity, complexities and the lack of appropriate tools may lead to cumbersome and sometimes challenging conversions between these formats. With **Bioconvert**, we cover a wide spectrum of format conversions in a single entry point. To do so, we design a simple framework that allows to use existing tools when available and to implement new conversions when they do not exist.

Bioconvert project has only recently started (at the end of 2017), nevertheless, thanks to a collaborative approach, there are already about 90 conversions available, including 40 different formats. Each conversion may have different implementations leading to about 120 unique methods.

Bioconvert is available on github at https://github.com/bioconvert/bioconvert and from pypi website. The project follows modern software development and good practices including openness, testing, continuous integration and automatic online documentation. Documentation is available at: http://bioconvert.readthedocs.io . In addition to a standalone version available from source or via existing bio-containers (on bioconda or as Singularity image), we also plan to provide an online version where users can easily convert their data without having to install the software locally.

http://bioconvert.readthedocs.io/en/dev/_images/conversion.png

Acknowledgements

We thank all contributors to the bioconvert project namely: Anne Biton, Bryan Brancotte, Yoann Dufresne, Kenzo-Hugo Hillion, Etienne Kornobis, Pierre Lechat, Rachel Legendre, Frédéric Lemoine, Blaise Li, Nicolas Maillet, Amandine Perrin, Rachel Torchet, Nicolas Traut, Anna Zhukova.

References

1. Documentation: http://bioconvert.readthedocs.io/
2. Github : https://github.com/bioconvert/bioconvert

Bioinformatic characterization of the role of TRIP12 in pancreatic adenocarcinoma

Lobna OUESLATI, Jérôme TORRISANI and Vera PANCALDI

Centre de recherche en Cancérologie de Toulouse (CRCT, UMR1037 Inserm / Université Toulouse III Paul Sabatier, 2 Avenue Hubert Curien, 31037, Toulouse, France)

Corresponding Author: lobna.oueslati@inserm.fr, vera.pancaldi@inserm.fr

1. Introduction

In 2020 pancreatic ductal adenocarcinoma (PDAC) will be the first cause of cancer related death in France, mostly due to late detection and the lack of effective treatment options for most patients. PDAC is the most common type of pancreatic cancer, affecting the duct-like epithelium by interspersion with less differentiated epithelial cells contained within a sea of proliferative stroma. During the last 40 years, no major advancements were made in PDAC prognosis as most therapies fail in controlling this cancer's progression [1]. This unmet need pathology warrants a deeper investigation of factors influencing incidence and resistance to therapy of PDAC. Beyond its known role in DNA damage response, TRIP12 was recently identified in the lab as a promising new PDAC target as it is involved in the maintenance of the acinar phenotype in healthy pancreas.

2. Results

One of the targets of TRIP12 is ptf1a, an important regulator of pancreatic differentiation [4]. Additionally. TRIP12 was found to associate to chromatin in large foci visible through immunofluorescence microscopy and over-expression of the protein (specifically the IDR region) produced a lethal aggregation of chromatin. The presence of TRIP12 all through the cell cycle and not just limited to S phase suggest that it might have unexplored roles in non-dividing cells. A series of experiments were conducted to alter TRIP12 expression using silencing RNAs generating multi-omics data in TRIP12 knock-down and WT HeLa cells and in the MIA-PaCa-2 PDAC cell line.

We used a protein-protein physical interaction network (thebiogrid.org) to project differential expression between WT and knock-down conditions and identified network modules that are co-ordinately regulated in the knock-out. Functional enrichment analyses confirmed the role of TRIP12 in chromatin organization (known interactors involve members of the Polycomb and SWI/SNF complex). Further, we intersected genes differentially regulated in TRIP12 knock-downs in MIA-Pa-Ca-2 and HeLa cell lines, finding a consistent overlap indicating possible conserved functions of TRIP12 in these two systems. Exploration of the members of these regulated pathways suggests possible mechanisms for TRIP12 to contribute directly or indirectly to the PDAC phenotype with roles beyond those already known.

References

1. Irfana Muqbil et al. Systems and Network Pharmacology Strategies for Pancreatic Ductal Adenocarcinoma Therapy: A Resource Review in Molecular Diagnostics and Treatment of Pancreatic Cancer: Systems and Network Biology Ap-proaches, 405-425, 2014.
2. N.C. Barmswig et al. Identification of new TRIP12 variants and detailed clinical evaluation of individuals with non-syndromic intellectual disability with or without autism. *Human Genetics* 136(2): 179-192, 2017.
3. X. Liu et al. Trip12 is an E3 ubiquitin ligase for USP7/HAUSP involved in the DNA damage response. *FEBS Letters* 590(23): 4213-4222, 2016.
4. Naima Hanoun et al. The E3 Ubiquitin Ligase Thyroid Hormone Receptor-interacting Protein 12 Targets Pancreas Transcription Factor 1a for Proteasomal Degradation *J Biol Chem.* 289(51): 35593–35604, 2014.

Biomarkers for neurodegenerative diseases

Céline LE BÉGUEC[1], Vincent ANQUETIL[2,3], Ivan MOSZER[2], Isabelle LE BER[2,3,4], Olivier COLLIOT[2,3], Pierre PETERLONGO[1], Dominique LAVENIER[1]

[1] University Rennes, Inria, CNRS, IRISA, F-35042 Rennes, France
[2] Inserm U 1127, CNRS UMR 7225, Institut du Cerveau et de la Moelle Épinière, ICM, University Sorbonne, Paris, France
[3] Assistance Publique - Hôpitaux de Paris, Hôpital Pitié-Salpêtrière, Paris, France
[4] Reference Center for rare or early dementias, IM2A, AP-HP, Paris, France

Corresponding Author: celine.le-beguec@inria.fr

In Europe it has been shown that one in eight people will be affected by diseases of the nervous system. This figure may increase in the coming years as the population ages. Alzheimer's (AD) or Parkinson's diseases (PD) affect respectively nearly 900,000 and 150,000 people in France, and 30 million and 6.3 million in the world. These diseases are increasingly well detected but cannot be cured. A slowdown can only be envisaged. Causes (genetic, environment...) of other, rare diseases, such as Fronto-Temporal Dementias (FTD, 4 to 10 cases per 100,000 inhabitants) or Amyotrophic Lateral Sclerosis (ALS, 5 to 7 cases per 100,000 inhabitants) have been discovered. So far the diagnosis of these diseases remains difficult as the overlap AD is high, making them difficult to differenciate at early stages. Therefore, biomarkers are needed. The Neuromarkers project (Inria Project Labs) aims in i) discovering new SNPs related to specific neurodegenerative diseases, and ii) identifying patterns of SNPs or genes according to their expression.

A new approach based on the DiscoSnp++ [1] tool developed by the GenScale team (Inria/IRISA, Rennes) will be used to discover new SNPs from various panels of patients. Thanks to the absence of mapping and to its indexing data-structure, is faster and requires less memory than classical SNP discovery tools. We will challenge and validate DiscoSnp++ with already analyzed clinical data from Whole Exome Sequencing (WES).

A second part of the project consists to correlate neurodegenerative diseases to the smallest number of significant patterns of SNPs. The goal is to find groups of SNPs that could explain the difference between cases and controls, and between several neurodegenerative disorders. A software based on pattern mining techniques and developed by the GenScale and Lacodam teams (Inria/IRISA, Rennes) will be used for that purpose. From a large number of SNPs and two cohorts of patients (case/control) the software, called SSDPS (Statistically Significant Discriminative Patterns Search) [2], extracts patterns of SNPs. In addition, using exhaustive miRNA expression high-throughput data, we will use gene expression to discriminate gene groups that can explain the different phenotypes observed. This part would allow additional validation of the method.

The Neuromakers project gathers researchers from several Inria teams and clinician and biologists from the Brain and Spine Institute (ICM, Paris).

Acknowledgements

This work was supported by the IPL – Neurmarkers.

References

[1] R. Uricaru *et al.*, “Reference-free detection of isolated SNPs.,” *Nucleic Acids Res.*, vol. 43, no. 2, p. e11, 2015.

[2] H. Son Pham, “Novel Pattern Mining Techniques for Genome-wide Association Studies,” *HAL*, 2017.

Burrowing functional and immunogenetic information through the 1000 Genomes Project with Ferret v.3.0

Rokhaya BA[1,2,3], Nicolas VINCE[1,2], Estelle GEFFARD[1,2], Dorian MALGUID[3],
Marie LANZA[3], Pierre-Antoine GOURRAUD[1,2] and Sophie LIMOU[1,2,3]

[1] ATIP-Avenir, Centre de Recherche en Transplantation et Immunologie UMR 1064, INSERM, Université de Nantes, Nantes, France
[2] Institut de Transplantation Urologie Néphrologie (ITUN), CHU Nantes, Nantes, France
[3] Ecole Centrale de Nantes, Nantes, France

Corresponding Author: sophie.limou@univ-nantes.fr

The 1000 Genomes project[1] (1KG) aims to catalog the common human genetic variations, including high-resolution HLA typing, in worldwide reference populations. We developed "Ferret"[2], a user-friendly Java tool, to ease the access of the community to the large and complex 1KG genomic data files. Ferret provides unique features including multiple input formats (locus, gene or SNP), fast extraction of individual genotype data, allele frequencies computation, and standard output formats.
Here, we present a new Ferret release (version 3.0) offering functional annotations designed to enable users to prioritize genetic variants and interpret genetic associations based on biological and immunogenetic contexts. Ferret v3.0 now grants access to 1) high-resolution HLA alleles[3] (HLA-A, -B, -C, -DR and -DQ), 2) basic functional annotations (gene name, variant location, amino acid consequences), and 3) advanced functional annotations (prediction on gene expression regulation and protein function). Advanced functional annotations include results from SIFT and PolyPhen, two bioinformatic tools predicting the impact of amino acid substitutions on protein structure and function, and from RegulomeDB, a database compiling DNA features and regulatory elements in non-coding regions of the human genome. Adding functional annotations initially increased Ferret runtime by 4-5 during our preliminary tests, but parallelizing tasks eventually limited the impact of annotation on runtime speed (x1.2). Retrieving HLA alleles through Ferret will offer a unique opportunity to explore this major and complex locus in genetic population studies and to go beyond simple SNP associations in genetic association studies.
In conclusion, Ferret delivers a straightforward way, even for clinicians and biologists, to manipulate, explore, and exploit 1KG data while providing biological and functional annotations. This tool could therefore empower the community to leverage the 1KG data and gain the most complete information on genes/loci of interest. Ferret is publicly available at: http://limousophie35.github.io/Ferret/.

References

1. 1000 Genomes Project Consortium. "A global reference for human genetic variation." *Nature* 526.7571 (2015): 68.
2. Limou, Sophie, Andrew M. Taverner, and Cheryl A. Winkler. "Ferret: a user-friendly Java tool to extract data from the 1000 Genomes Project." *Bioinformatics* 32.14 (2016): 2224-2226.
3. Gourraud, Pierre-Antoine, et al. "HLA diversity in the 1000 genomes dataset." *PloS one* 9.7 (2014): e97282.

CADBIOM – Un logiciel pour l'identification de contrôleurs dans des réseaux de signalisation et de régulation extraits de Pathway Commons

Pierre VIGNET[3, 1], Jean COQUET[2], Nathalie THERET[3] and Anne SIEGEL[1]

[1] Univ Rennes, Inserm, EHESP, Irset -UMR_S1085, F-35000 Rennes, France

[2] Stanford University, Biomedical Informatics Research, CA 94305, Stanford, États-Unis

[3] Univ Rennes, Inria, CNRS, IRISA, F-35000, Rennes, France

Corresponding Author: pierre.vignet@irisa.fr

La modélisation de la dynamique discrète des réseaux de signalisation s'appuie sur divers formalismes faisant appel à des règles, à des modèles logiques ou encore à des automates cellulaires (Biocham[1], Pint[2], Caspo[3], Kappa[4]). Toutefois, ces approches ne peuvent s'appliquer à des modèles à grande échelle et leur construction est basée sur la sélection *a priori* des molécules du modèle. La construction automatique sans *a priori* de grands modèles implique l'utilisation de sources de connaissances standardisées et le langage BioPAX[5] remplit pleinement cette fonction. C'est sous ce format que la base de données Pathway Commons[6] centralise les connaissances issues de 23 autres bases de données, soit la description de 4700 voies de signalisation et de régulation, représentant 2,3 millions d'interactions. L'enjeu consiste aujourd'hui à exploiter ces sources de données pour créer et analyser sans *a priori* des modèles dynamiques reposant sur des réseaux de signalisation et régulation.

Cadbiom a été conçu comme un logiciel d'analyse de systèmes biologiques[7] permettant de formaliser les réactions biologiques sous forme de transition gardées et d'explorer un réseau à l'aide de requêtes de causalité. Ce logiciel permet d'une part de construire de grands modèles à partir de connaissances formalisées en langage BioPAX, et d'autre part d'analyser ces réseaux pour identifier des régulateurs (gènes, protéines) de cibles d'intérêt. L'usage de Cadbiom se résume en quatre modules. Un premier module permet d'interroger les bases de données pour extraire les connaissances BioPAX et formaliser un modèle Cadbiom. Un deuxième module est dédié à la construction de requêtes d'intérêt à partir d'une liste d'identifiants de biomolécules (HGNC/Uniprot). Le troisième module explore le modèle et permet d'identifier des ensembles de biomolécules/contrôleurs du système menant à l'activation des entités mentionnées dans les requêtes. Le quatrième module permet d'interpréter les solutions et les trajectoires associées à l'aide d'outils de visualisation (*heat maps*, graphes).

Cadbiom permet d'intégrer, exporter et explorer de grands réseaux de signalisation. À titre d'exemple, les requêtes sur des gènes impliqués dans la transition épithélio-mésenchymateuse ont permis de mettre en évidence des mécanismes de régulation propres à la coactivation de gènes.

Cadbiom est écrit en Python; la documentation est accessible sur http://cadbiom.genouest.org et les modules sont distribués sous licence libre GNU GPL sur PyPI (*Python Package Index*).

Références

1. L. Calzone, F. Fages, et S. Soliman, BIOCHAM: an environment for modeling biological systems and formalizing experimental knowledge, *Bioinformatics*, vol. 22, n° 14, p. 1805-1807, 2006.
2. L. Paulevé, Pint: A Static Analyzer for Transient Dynamics of Qualitative Networks with IPython Interface, in *Computational Methods in Systems Biology*, p. 309-316, 2017.
3. S. Videla, J. Saez-Rodriguez, C. Guziolowski, et A. Siegel, caspo: a toolbox for automated reasoning on the response of logical signaling networks families, *Bioinformatics*, vol. 33, n° 6, p. 947-950, 2017.
4. P. Boutillier *et al.*, The Kappa platform for rule-based modeling, *Bioinformatics*, vol. 34, n° 13, p. i583-i592, 2018.
5. E. Demir *et al.*, BioPAX – A community standard for pathway data sharing, *Nat Biotechnol*, vol. 28, n° 9, p. 935-942, 2010.
6. E. G. Cerami *et al.*, Pathway Commons, a web resource for biological pathway data, *Nucleic Acids Res*, vol. 39, n° Database issue, p. D685-D690, 2011.
7. G. Andrieux, M. Le Borgne, et N. Théret, An integrative modeling framework reveals plasticity of TGF-β signaling, *BMC Syst Biol*, vol. 8, p. 30, 2014.

Can we detect DNA methylation with Oxford Nanopore reads ?

Paul Terzian[1], Céline Vandecasteele[2], Alice Guidot[3], Ludovic Legrand[3], Christine Gaspin[2] , Denis Milan[2], Cécile Donnadieu[2], Carole Iampietro[2] and Christophe Klopp[1]

[1] MIAT, PF Bioinfo GenoToul, Université de Toulouse, INRA, Chemin de Borde Rouge, 31320 Castanet-Tolosan, France

[2] INRA, US 1426, GeT-PlaGe, Genotoul, 31320 Castanet-Tolosan, France

[3] LIPM INRA/CNRS, chemin de Borde Rouge, 31320 Castanet-Tolosan,France

Corresponding Author: paul.terzian@inra.fr

Long read sequencing technologies enable to call nucleotides and detect methylated sites with the same signal. Oxford Nanopore Technologies (ONT) made this possible using electrical current variations produced by a DNA strand passing through a pore. Despite their great potential, ONT sequencers and the corresponding methylation detection algorithms are still very recent and therefore need to be validated. Our study aims at comparing ONT methylome analysis methods in order to define the conditions in which they work best.

Most of the available detection software packages are model-based, implying using existing models or training your own. The model training algorithms use from Hidden Markov Models (mcaller[1], nanopolish [2]) for the oldest ones to deep learning (DeepSignal [3]) for the latest. We first tested the Tombo package [4], ONT official methylation detection framework. This python package provides 5 methylation detection models : generic models for 5mC (5-Methylcytosine) and 6mA modifications (*6*-Methyladenine), dam and dcm motif specific models (found in E. coli [5]) and a CpG (found in human).

Our preliminary analysis focused on *Ralstonia solanacearum,* a plant pathogenic bacterium studied in the EPI-PATH ANR project. PacBio long-reads had already been obtained and analyzed for methylation detection [6], enabling us to compare both technologies. ONT reads showed a statistical difference in methylation between wild type and GTWWAC motif methyltransferase knock-downed strains, as expected. We compared these detection results with those obtained with PacBio reads and observed shared tendencies but poor genomic position fraction correlation. These first results motivated us to train a species specific model and compare it to the supplied generalist model.

Acknowledgements

This study is part of the SeqOccin project (https://get.genotoul.fr/seqoccin/) which aims at gaining expertise in several long reads sequencing technology applications. Get-Plage and Genotoul Bioinfo run this 3 years long project helped by many interested scientists.

References

[1] McIntyre, A, B., Alexander N., Grigorev K., Bezdan, D., Sichtig, H., Chiu, C. Y., & Mason, C. E. (2019). Single-molecule sequencing detection of N6-methyladenine in microbial reference materials. *Nature communications, 10*(1), 579.

[2] Loman NJ, Quick J, Simpson JT. 2015. A complete bacterial genome assembled de novo using only nanopore sequencing data. *Nature methods* 12 : 733-735.

[3] Ni, P., Huang, N., Luo, F., & Wang, J. (2018). DeepSignal: detecting DNA methylation state from Nanopore sequencing reads using deep-learning. *BioRxiv*, 385849.

[4] Stoiber, M. H., Quick, J., Egan, R., Lee, J. E., Celniker, S. E., Neely, R., ... & Brown, J. B. (2017). De novo identification of DNA modifications enabled by genome-guided nanopore signal processing. *BioRxiv*, 094672.

[5] Palmer, B. R., & Marnius, M. G. (1994). The dam and dcm strains of Escherichia coli – a review. *Gene*, *143*(1), 1-12.

[6] Erill, I., Puigvert, M., Legrand, L., Guarischi-Sousa, R., Vandecasteele, C., Stubal, J. C., … & Valls, M. (2017). Comparative analysis of Ralstonia solanacearum methylomes. *Frontiers in plant science, 8*, 504.

Caractérisation de CNV (variants de nombre de copies) à partir de données de séquences exoniques simulées

Quentin MIAGOUX[1], Maëva VEYSSIERE[1], Anna NIARAKIS[1], Elisabeth PETIT-TEIXEIRA[1] et Valérie CHAUDRU[1]

[1] Genhotel-EA3886, Univ Evry, Université Paris Saclay, 2, rue Gaston Crémieux, 91000, EVRY-GENOPOLE cedex, France

Corresponding Author: quentin.miagoux@univ-evry.fr

1. Introduction

Les variations de nombre de copies (CNV), conduisant à des gains ou des pertes de segments chromosomiques de taille supérieure à 1 kb, composent environ 12% du génome humain [1]. Plusieurs études ont montré que ces CNV pouvaient être associés à des maladies multifactorielles, comme la Polyarthrite Rhumatoïde (PR) [2,3]. Les études réalisées dans la PR étant principalement orientées vers des CNV candidats, notre objectif est d'identifier de nouveaux CNV associés à cette pathologie à partir de données de séquences exome-entier pour des individus appartenant à 9 familles françaises à cas multiples de PR. Au total, 30 individus (19 atteints/11 non atteints) ont été séquencés. Préalablement à l'analyse de ces données, nous avons testé, sur des données simulées, deux outils de détection de CNV à partir de données de séquence d'exome, Codex2 et ExomeDepth [4,5]. Ces outils semblent avoir la meilleure sensibilité parmi les outils existants pour trouver les régions CNV [6] mais on ne connait pas leur sensibilité pour caractériser les CNV par individu. Or, une mauvaise attribution de CNV par individu peut conduire à des résultats d'association erronés.

2. Matériel et méthodes

Nous avons simulé, pour 30 individus et une profondeur de lecture de 100X, 64 CNV (33 duplications, 31 délétions), de tailles et fréquences variables, répartis sur 3 chromosomes (1, 17 et 22). Ces simulations ont été réalisées à l'aide des packages RSVSim [7] et Wessim2 [8].

3. Résultats et discussion

Sur les 64 régions CNV simulées, le pourcentage de CNV détectés par rapport aux vrais CNV était de 64% avec Codex2 et de 58% avec ExomeDepth, ces résultats étant cohérents avec les données de la littérature. Les CNV qui n'ont pas été détectés avec les 2 outils étaient plutôt des CNV de petite taille (< 5 kb) et/ou incluant uniquement une partie d'un exon. En se restreignant aux régions correctement identifiées à l'étape précédente, 50% des CNV introduits étaient bien caractérisés chez les individus avec ExomeDepth et 72% avec Codex2. Ces résultats varient cependant en fonction de la fréquence du CNV : plus il est rare, mieux il sera identifié chez les individus. Ce travail d'analyse de données simulées se poursuit en testant d'autres outils de détections de CNV qui tiennent également compte des séquences hors cibles.

Références

1. Redon R, Ishikawa S, Fitch KR, et al. Global variation in copy number in the human genome. *Nature*. 2006;444(7118):444–454.
2. Chen J, Huang F, Liu M, Duan X, Xiang Z. Genetic polymorphism of glutathione S-transferase T1 and the risk of rheumatoid arthritis: a meta-analysis. Clin Exp Rheumatol. 2012 Sep-Oct;30(5):741-7. Review.
3. Schaschl H, Aitman TJ, Vyse TJ. Copy number variation in the human genome and its implication in autoimmunity. Clin Exp Immunol. 2009;156(1):12–16.
4. Jiang Y, Wang R, Urrutia E, et al. CODEX2: full-spectrum copy number variation detection by high-throughput DNA sequencing. *Genome Biol*. 2018;19(1):202.
5. Plagnol V, Curtis J, Epstein M, et al. A robust model for read count data in exome sequencing experiments and implications for copy number variant calling. *Bioinformatics*.
6. Sadedin SP, Ellis JA, Masters SL, Oshlack A. Ximmer: a system for improving accuracy and consistency of CNV calling from exome data. *Gigascience*. 2018;7(10):giy112. Published 2018 Sep 6. doi:10.1093/gigascience/giy112
7. Bartenhagen C (2018). *RSVSim: RSVSim: an R/Bioconductor package for the simulation of structural variations*.
8. Kim S, Jeong K, Bafna V. Wessim: a whole-exome sequencing simulator based on in silico exome capture. *Bioinformatics*. 2013;29(8):1076–1077.

CD4 T cell reprogramming in brain-injured patients

Alice MOLLE[1,2], Cynthia FOURGEUX[1,2], Tanguy CHAUMETTE[3], Jeremie POSCHMANN[1,2] and Antoine ROQUILLY[3]

[1] Centre de Recherche en Transplantation et Immunologie UMR 1064, INSERM, Université de Nantes,30 Boulevard Jean Monnet, 44093, Nantes, France

[2] Unité de Transplantation Urologie Néphrologie (ITUN), 30 Boulevard Jean Monnet, 44093, Nantes, France

[3] Intensive Care Unit, Hotel Dieu, Place Ricordeau, 44903, Nantes France

Corresponding Author: Jeremie.Poschmann@univ-nantes.fr

A frequent effect observed in traumatized patients is a Systemic Inflammatory Response Syndrome (SIRS) characterized by an excessive release of inflammatory cytokines that activate innate and acquired immunity. Paradoxally, this inflammation is followed by an anti-inflammatory phase that lasts in time [1]. This immunosuppression is accompanied in patients by an increase of secondary infections which are a main cause of mortality and morbidity in intensive care units. One of these secondary infections is the reactivation of herpes simplex virus (HSV) when immunity, primarly the memory T cells, fails to control replication [2].

Previous results have shown an association between early lymphopenia and survival in brain-injured patients. While cytokine production by CD8 T cells remains stable, the percentages of IFN-g^+ CD4 T cells decreased from day 1 followed by incomplete recovery at 6 months. In addition, it was found that other cytokines and transcriptional factors remained changed even at 6 months. The cellular micro-environment has been shown to maintain reprogramming by cytokines released, notably circulation monocytes.

This study aims to investigate the intrinsic re-programming of CD4 T cells after trauma by examining the differential level of histone acetylation among controls and patients with or without HSV reactivation. An epigenetic analysis of the H3K27ac profiles have highlighted distinct epigenetic signatures between healthy volunteers and brain-injured patients that could be detected since day 1. We also found a list of differentially regulated genes and associated gene ontology. This list contained many transcription factors, some of which were found at the level of enriched binding motifs and are being validated. All these experiments will allow us to better understand the mechanisms involved in this re-programming and to adjust the treatments.

References

1. ABalk RA. Systemic inflammatory response syndrome (SIRS): where did it come from and is it still relevant today?. *Virulence*. 2014

2. Sundar KM, Sires M. Sepsis induced immunosuppression: Implications for secondary infections and complications. *Indian J Crit Care Med*. 2013

cDNA length improvement is essential to allow better isoform characterization for long read RNA sequencing

Sophie LEMOINE[1], Ammara MOHAMMAD[1,2], Corinne BLUGEON[1], Bérengère LAFFAY[1,3], Laurent JOURDREN[1]

[1] Institut de Biologie de l'ENS (IBENS), Département de biologie, École normale supérieure, CNRS, INSERM, Université PSL, 75005 Paris, France

[2] INSERM U1127, CNRS UMR 7225, Sorbonne Universités, UPMC Univ Paris 06 UMRS 1127, Institut du Cerveau et de la Moelle épinière, ICM, Paris, France

[3] Master Bioinformatique, Normandie Université, UNIROUEN (UNIROUEN), Université de Rouen Normandie, France

Corresponding Author: slemoine@biologie.ens.fr

Estimation of transcript isoform is a real challenge with short read sequencing. With Oxford Nanopore Technologies (ONT), our aim is to sequence full-length cDNA in order to directly access transcript isoforms.

We have successfully validated analysis of differential expressed targets on a mouse model of myelination blockage (Egr2 knock-out)[1] from 100ng of RNA following the standard ONT protocol. The mean length of our aligned reads was 1.2kb, which is lower than the estimated 2kb mean length of the mouse validated and modeled messenger RNAs from Ensembl[2] and even worse if we only consider the TSL1 tagged mouse transcripts (2.6kb). To improve our protocol and lower the amount of input RNA, we looked for other cDNA synthesis options. We therefore used the SmartSeq technology (Clontech/TakaraBio) to synthesize full-length cDNA from only 10ng of total RNA. Sequencing adapters were added using the ONT ligation 1D protocol. The cDNAs obtained were barcoded in order to sequence multiple samples on a single MinION run and allow differential expression analyses. The sequenced reads were mapped on the mouse genome using Minimap2[3] and differential expression analysis was performed using DESeq2[4] on the Eoulsan[5] pipeline.

We compared the sequences from SmartSeq to the one from ONT standard protocol. We found that SmartSeq technology allowed us to sequence much longer cDNAs. The mean length of the reads was then about 2.6kb and the small reads that were the majority of the population with ONT was nearly eradicated. We were able to detect more differentially expressed targets with less input material. The supplementary targets detected were longer compared than the ONT protocol ones. Overall, the optimized protocol globally achieved a better 5'-3' coverage for transcripts and not surprisingly, for longer than 2kb transcripts. If SmartSeq technology does not ensure you have full-length cDNAs, it proves it can be a reliable option for cDNA sequencing on the MinION and improve isoform annotation and quantification using pipelines such as FLAIR[6].

References

1. Topilko P. *et al*., 1994. Nature 371.796-799
2. Daniel R. *et al* Ensembl 2018. doi:10.1093/nar/gkx1098
3. Li, H. (2018). Minimap2: pairwise alignment for nucleotide sequences. Bioinformatics, 34:3094-3100. doi:10.1093/bioinformatics/bty191
4. Love MI, Huber W and Anders S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biology 2014 15:550. doi:10.1186/s13059-014-0550-8
5. Laurent Jourdren, Maria Bernard, Marie-Agnès Dillies and Stéphane Le Crom. Bioinformatics (2012) 28 (11): 1542-1543. doi:10.1093/bioinformatics/bts165
6. Rachael E Workman *et al*; Nanopore native RNA sequencing of a human poly(A) transcriptome; bioRxiv 459529; doi:10.1101/459529

Characterization of Hepatitis B Virus genomes identified by viral capture in Hepatocellular Carcinomas from European and African patients

Camille PENEAU[1,2], Sandrine IMBEAUD[1,2], Tiziana LA BELLA[1,2], Iadh MAMI[1,2], Jessica ZUCMAN-ROSSI[1,2,3]

[1] Centre de Recherche des Cordeliers, Sorbonne Universités, Inserm, F-75006 Paris, France
[2] Functional Genomics of Solid Tumors, Université de Paris, Université Paris 13, Labex Immuno-Oncology, équipe labellisée Ligue Contre le Cancer, F-75000 Paris, France
[3] European Hospital Georges Pompidou, AP-HP, F-75015, Paris, France

Corresponding Author: camille.peneau@inserm.fr | jessica.zucman-rossi@inserm.fr

Abstract: Viruses are common causes of cancer in humans and, among them, the Hepatitis B virus (HBV) was identified as the leading risk factor for hepatocellular carcinoma (HCC) occurrence, the third cause of cancer death worldwide. HBV-related HCC develop not only in the setting of cirrhosis but also in normal liver, underlying that the virus has its own oncogenic properties. HBV is a DNA virus that could integrate in human DNA and promote cell transformation by insertional mutagenesis. As HBV integrations occur early during infection, it is a key factor reflected in the genetic landscape of HCC. Therefore comparing the insertions occurring in normal hepatocytes and in tumor cells may enable to identify new genomic defects associated with tumor development. The recent development of next generation sequencing has given the opportunity to characterize more precisely the role of HBV insertion as a cancer driver alteration. Cancer-related genes such as *TERT*, *CCNE1* and *MLL4* have been identified as recurrently targeted by HBV insertions in HCC, with specific biological consequences and clinical outcomes. However, up to now, such identifications of HBV insertions in HCC have been mainly performed in Asian populations. Our project aimed to characterize HBV-related insertional mutagenesis in tumor and non-tumor liver tissues, in a large cohort of HCC from patients with European or African origin.

We performed viral capture and next-generation sequencing on 220 HCC and their normal liver counterparts from an in-house series of 180 HBV-positive patients. The capture was optimized to fragment the genomic DNA to 1kb in length and to sequence 48 samples in one run. We set up a new pipeline of analysis in order to characterize precisely not only the integration sites in each tissue but also the integrated viral sequences. By normalizing the sequencing coverage at the breakpoints of integration, we estimated the clonality of each insertion to highlight the role of clonal and sub-clonal integrations in HCC development and establish a timing of the integration events. 74% of the HBV-integrated tumors have more than one clonal insertion site, suggesting that integration occurs multiple times in the same cellular clone and underlying the importance of understanding the chronology. For each insertion locus, we extracted the paired-end reads mapping on both sides of the integration breakpoint and the chimeric reads covering the junction, in order to reconstruct the different integrated viral sequences present within each sample. This in-silico reconstruction of integrated sequences in tumors revealed the existence of frequent structural rearrangements in the viral sequence (inversions, deletions, duplications) as in the human genome around the integration breakpoints (translocations, large deletions, amplifications). These results suggest that HBV-related insertional mutagenesis may derive from or result in an increased chromosomal instability, altering cancer-related genes, which may be crucial to trigger the initial clonal expansion. But the consequences of the rearrangements observed remain to be investigated to better understand the whole mechanism of HBV insertion.

This study provides a global view of the landscape of HBV integrations in European and African populations, by characterizing the different viral forms and sequences in tumors and non-tumor liver tissues. This unique dataset can now be correlated with the genetic and clinical features of the patients to identify which alterations are early or late during the process of hepato-carcinogenesis, to decipher the viral and human genomic context in which these insertions occur, and to explore the impact of anti-viral treatments on HCC development.

checkMyIndex: a web-based R/Shiny interface for choosing compatible sequencing indexes

Hugo Varet*[1,2] and Jean-Yves Coppée[1]

[1]Biomics Pole - C2RT, Institut Pasteur, Paris, France – Institut Pasteur de Paris – France
[2]Hub Bioinformatique et Biostatistique - Bioinformatics and Biostatistics HUB – Institut Pasteur [Paris], Centre National de la Recherche Scientifique : USR3756 – France

Résumé

When sequencing several libraries simultaneously, the selection of compatible combinations of indexes is critical for ensuring that the sequencer will be able to decipher the short, sample-specific barcodes added to each fragment. However, researchers have few tools to help them choose optimal indexes. Here, we present checkMyIndex [1], an online R/Shiny application that facilitates the selection of the right indexes as a function of the experimental constraints.
We used the popular R package shiny to develop a user-friendly application (available free of charge at checkmyindex.pasteur.fr) dedicated to searching for compatible combinations of indexes. Our application supports both single- and dual-indexing and is compatible with the various chemistries used by the Illumina HiSeq, MiSeq, NextSeq and iSeq devices.

In practice, the user only needs to provide his/her available indexes as a simple, two-column, tab-separated text format file. The first column contains the index identifiers, and the second contains the corresponding short sequences. Next, several constraints have to be defined via the interface: the total number of samples, the multiplexing rate (i.e. the number of samples per pool/lane) and whether the same index or the same combination of indexes can be used several times.

Solutions are returned quickly (in a few seconds) by the underlying algorithm, and satisfy the constraints imposed by the user. Moreover, the structure of the R code will allow new features to be added, e.g. when new chemistries are developed.
Hugo Varet, Jean-Yves Coppée; checkMyIndex: a web-based R/Shiny interface for choosing compatible sequencing indexes, Bioinformatics, 2018, bty706, https://doi.org/10.1093/bioinformatics/bty706

*Intervenant

ChIPuana: from raw data to epigenomic dynamics

Maëlle Daunesse[1], Rachel Legendre[2], Hugo Varet[2], Thomas Cokelaer[2], Claudia Chica[1*]

[1] Institut Pasteur, Biostatistics and Bioinformatics Hub, C3BI
[2] Institut Pasteur, BIOMICS
[*] cchica@pasteur.fr

We present ChIPuana, a snakemake-based workflow for the analysis of epigenomic data (ChIPseq) from the raw fastq files to the differential analysis of transcription factor binding or histone modification marking. It streamlines critical steps like the quality assessment of the immunoprecipitation using the cross correlation and the replicate comparison for both narrow and broad peaks. For the differential analysis ChIPuana provides linear and non linear methods for normalisation between samples as well as conservative and stringent models for estimating the variance and testing the significance of the observed differences. We show examples of how various settings can allow users to improve the discriminative power of their comparisons depending on the dynamics of the epigenomic factor under study.

ChIPana is implemented in Sequana, a Python-based library that facilitates the installation of the dependencies and simplifies the modification and extension of the workflow. For the end-user the pipeline can be executed interactively via Sequanix, a user-friendly graphical interface. A complete report is produced at the end of the bioinformatic and the statistical part of the analysis to facilitate the interpretation of the results.

Keywords: Epigenome, ChIP-seq, Differential analysis, Dynamics of binding and marking, Snakemake

Chronic mood instability, cardiometabolic risk and functional impairment in bipolar patients: relevance of a multidimensional approach

Stevenn VOLANT[1], Aroldo A. DARGÉL[2] and Chantal HENRY[2]

[1] Institut Pasteur – Bioinformatics and Biostatistics Hub – C3BI, USR 3756 IP CNRS – Paris, France

[2] Institut Pasteur, Unité Perception et Mémoire, Paris, France

Corresponding Author: stevenn.volant@pasteur.fr

Remitted bipolar disorder (BD) patients frequently present with chronic mood instability likely associated with poor psychosocial and cognitive functioning and low-grade inflammation. However, clear, distinct clinical phenotypes among remitted BD patients have not yet been distinguished. Based on emotional hyper-reactivity and activation levels, we aim to characterize these patients with chronic mood instability, and to examine the heterogeneity of clinical phenotypes in terms of cardiometabolic risk, chronic inflammation, and functional impairment.

A total of 979 adult remitted BD patients evaluated in the French Network of BD. Using the Multidimensional Assessment of Thymic States (MATHYS) patients were assessed for mood and levels of activation, which are based on five dimensions of behavior: emotional reactivity, sensory-perception, psychomotor activity, motivation and cognition. Usual approach focuses on the total MATHYS, determining patient status by setting arbitral threshold. We developed multidimensional approaches, mainly based on machine learning algorithm, which enhance the characterization of BD patients.

Results: (i) Emotional hyper-reactivity and cardiometabolic risk: using Random Forest algorithms we found that patients with emotional hyper-reactivity (n=326) had significantly higher levels of low-grade inflammation, hypertension, fast glucose, as well as greater number of suicide attempts ($P<10^{-8}$) than those without emotional hyper-reactivity (n=308). This predictive model identified patients with emotional hyper-reactivity with 84.9% accuracy [1]. (ii) Activation levels, cardiovascular risk and functioning: Using a cluster-analytic approach, 979 remitted BD patients were grouped in four clusters according to their levels of activation. Clusters with increased activation levels had higher blood pressure and chronic inflammation compared to those with normal or hypo-activation levels ($P<0.0001$). Clusters with abnormal activation levels had poorer cognitive and psychosocial functioning compared to the normal activation cluster ($P<0.0001$) [2].

The assessment of dimensions of behavior alongside mood is clinically relevant, particularly for identifying BD patients at higher risk of cardiometabolic dysfunction, suicide, and functional impairment, to develop more individualized body-brain interventions.

References

1. Emotional hyper-reactivity and cardiometabolic risk in remitted bipolar patients: a machine learning approach. Dargél AA, Roussel F, Volant S, Etain B, Grant R, Azorin JM, M'Bailara K, Bellivier F, Bougerol T, Kahn JP, Roux P, Aubin V, Courtet P, Leboyer M; FACE- BD Collaborators, Kapczinski F, Henry C.. Acta Psychiatr Scand. 2018 May 15.

2. Activation Levels, Cardiovascular Risk, and Functional Impairment in Remitted Bipolar Patients: Clinical Relevance of a Dimensional Approach. Dargél A, A, Volant S, Saha S, Etain B, Grant R, Azorin J, -M, Gard S, Bellivier F, Bougerol T, Kahn J, -P, Roux P, Aubin V, Courtet P, Leboyer M, Scott J, Henry C: . Psychother Psychosom 2019;88:45-47

Classification of evolutionary trajectories of cognitive functions

Céline BOUGEL[1], Sébastien DEJEAN[2], Caroline Giulioli[1], Philippe SAINT-PIERRE[2], Nicolas SAVY[2], Sandrine ANDRIEU[1]

[1] Institut National de la Santé Et de la Recherche Médicale, UMR1027 - INSERM, 37 Allées Jule Guesde, 31000, Toulouse, France

[2] Institut de Mathématiques de Toulouse, UMR5219 – Université de Toulouse ; CNRS – UPS IMT, F-31062 Toulouse Cedex 9, France

Corresponding Author: sandrine.andrieu@univ-tlse3.fr

1. Introduction

While demographic aging is inevitable, preventive approaches may prove effective against the cognitive functions decline and Alzheimer's disease. Prevention trials implemented to slow down the decline in cognitive functions with age have yielded unconvincing results. The aim of our work was to cluster cognitive evolution profiles in the context of Alzheimer's disease prevention trials.

2. Material and Methods

The Multidomain Alzheimer Preventive Trial has been presented in [1]. The primary endpoint of this trial was a composite Z-score combining 4 cognitive tests. Data were collected longitudinally. To complete the primary analysis, we used 3 methods :

1°) the k-means for longitudinal data (kml), a method of clustering trajectories without distribution hypothesis. The aim is to divide the population into k homogeneous subgroups in terms of trajectories shape regardless of time. A kml extension allows clustering trajectories according to the shape of them [2], so this should lead to unstationary and homogeneous groups (outputs not shown).

2°) the hierarchical cluster analysis (HCA) [3] which is unsupervised and not required specifying the number of groups. We calculated rates of change between each visit for the composite Z-score to to refer to non-longitudinal data and apply the HCA. The aim is to create homogeneous subgroups with the most similar individuals (intra-class homogeneity) and well-defined groups (inter-class heterogeneity).

3°) the seriation which consisted in transforming and optimally reordering a data matrix in order to bring to light, via shades of colors [4], a structure of the data. We apply this method on same rates of change as HCA. Only lines were swapped to keep the measurements' chronology and we restricted us to three colors to caricaturize information. The objective is to have even more homogeneous groups than with the previous methods.

3. Results

All the methods described were tested on the MAPT data set, on complete follow-up (1143 subjects). We identified stationary profiles: with HCA in the first group (containing 93% of subjects, see Figure 1), with seriation in 12-36 months. We also observed evaluative profiles (seriation in 0-12 months) and all other groups with HCA (Fig.1). Groups' characteristics for each method were not statistically different, except for two baseline cognitive tests with seriation: Trail Making Test ($p < 0.01$) and Mini Mental Status Examination ($p = 0.03$) in the sense that constant group had the best average score (28.33 points).

4. Discussion/Conclusion

Exploratory methods, not widely mobilized in the biomedical field, were applied to cluster cognitive evolution profiles. The results led to observations that were difficult to interpret clinically. We conclude that, in prevention trial, population remains cognitively constant during the first three years. In addition, information provided is complex. Another approach as hidden Markov model must be considered for future.

Acknowledgements

This work was supported by Association Monégasque pour la recherche sur la maladie d'Alzheimer (AMPA) and Harmonie Mutuelle et la fondation de l'avenir.

References

1. ANDRIEU, Sandrine, GUYONNET, Sophie, COLEY, Nicola, *et al.*Effect of long-term omega 3 polyunsaturated fatty acid supplementation with or without multidomain intervention on cognitive function in elderly adults with memory complaints (MAPT): a randomised, placebo-controlled trial. *The Lancet Neurology*, 2017, vol. 16, no 5, p. 377-389.
2. GENOLINI, Christophe, ECOCHARD, René, BENGHEZAL, Mamoun, *et al.* kmlShape: An efficient method to cluster longitudinal data (Time-Series) according to their shapes. *Plos one*, 2016, vol. 11, no 6, p. e0150738.
3. ALLAM, Apparao, GUMPENY, R. Sridhar, *et al.* Analyzing microarray data of Alzheimer's using cluster analysis to identify the biomarker genes. *International journal of Alzheimer's disease*, 2012, vol. 2012
4. HAHSLER, Michael et HORNIK, Kurt. Dissimilarity plots: A visual exploration tool for partitional clustering. *Journal of Computational and Graphical Statistics*, 2011, vol. 20, no 2, p. 335-354..

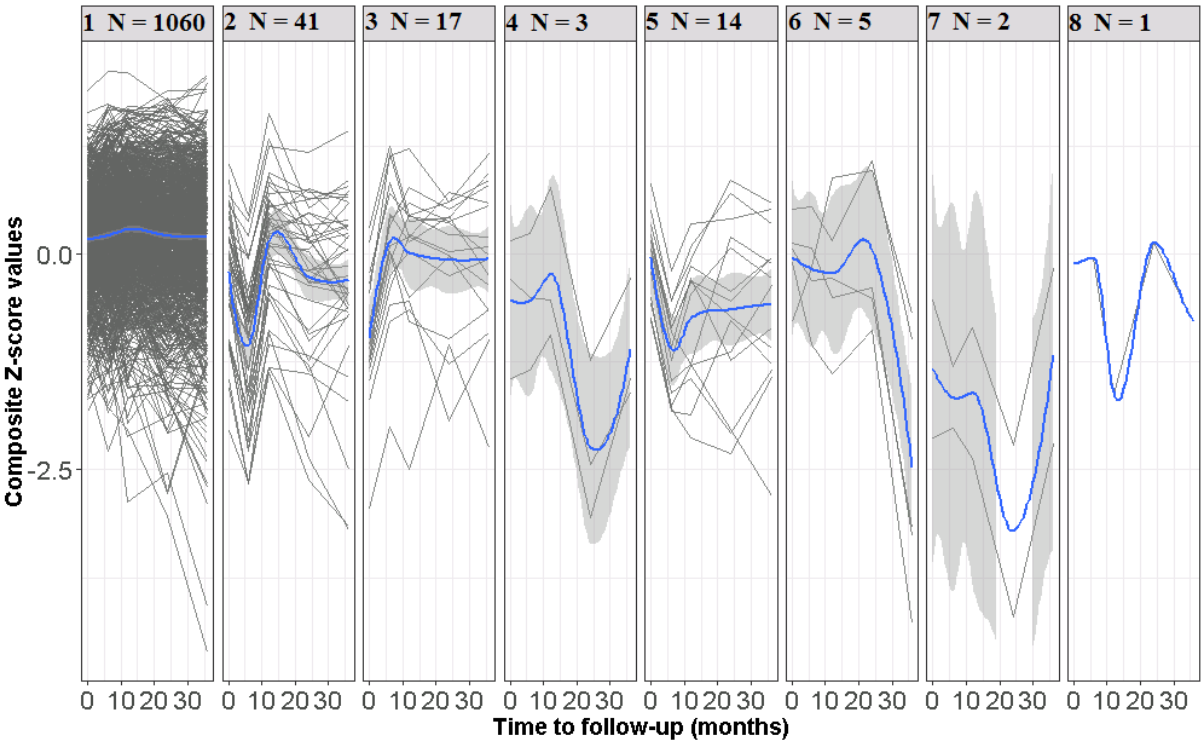


Figure 1- Representation of the different subgroups identified by the HCA method, MAPT data (N = 1143 subjects)

Co-activity networks reveal the structure of planktonic symbioses in the global ocean

Nils Giordano[1,2] and Samuel Chaffron[1,2]
[1] Université de Nantes, Centrale Nantes, CNRS UMR 6004, LS2N, F-44000 Nantes, France
[2] Research Federation (FR2022) Tara Oceans GO-SEE, Paris, France

Corresponding author: samuel.chaffron@univ-nantes.fr

Abstract

Marine microbes interact with their siblings and the environment, forming complex networks of connected metabolic and signaling pathways. Such communities play crucial ecological and biogeochemical roles on our planet, forming the basis of the marine food web, sustaining Earth's biogeochemical cycles in the oceans, and regulating climate. Limited by the fact that most microbes are difficult to isolate and cultivate in lab-controlled environments, we are just starting to grasp the complexity and diversity of their interactions. Today, large-scale environmental surveys of microbial communities (e.g. Tara Oceans expeditions [1]) gathered large volumes of meta-omic and contextual data that are enabling the reconstruction of genomes of uncultivated microbial species [2,3]. While classical co-occurrence analyses enable to predict potential interactions between these newly identified microbes [4], these approaches are inherently limited since true biotic interactions can hardly be disentangled from abiotic (environmental) effects.

Here, we propose a trait-based approach to enrich co-occurring information and uncover putative biotic interactions between marine bacterial organisms by directly inferring genomic and growth traits from meta-omics data. New methods have emerged to infer bacterial replication rates based on differential coverage in a metagenomic sample [5,6]. Available metatranscriptomic data also grant access to the expression of bacterial genomes in their environment. Across samples, these co-growth and co-expression signals can thus be exploited to reveal interactions between specific microbes and link their activities to the environmental context. In addition, we can use the functional content of these co-active genomes to predict their potential dependencies, in particular if they deviate from general scaling laws that govern the functional content of lab-cultivated microbial organisms [7]. Inferring and combining (meta-)genomic traits in a global framework can help to identify consortia of marine microbes and pave the way towards the functional understanding and the metabolic modeling of their interactions.

Acknowledgements

This work was supported by Université Bretagne Loire (UBL) and Recherche-Formation-Innovation (RFI).

References

[1] Shinichi Sunagawa, Luis Pedro Coelho, Samuel Chaffron, Jens Roat Kultima, Karine Labadie, et al. Structure and function of the global ocean microbiome. *Science*, 348(6237):1261359, May 2015.

[2] Benjamin J. Tully, Elaina D. Graham, and John F. Heidelberg. The reconstruction of 2,631 draft metagenome-assembled genomes from the global oceans. *Scientific Data*, 5:170203, January 2018.

[3] Tom O. Delmont, Christopher Quince, Alon Shaiber, Özcan C. Esen, Sonny TM Lee, et al. Nitrogen-fixing populations of Planctomycetes and Proteobacteria are abundant in surface ocean metagenomes. *Nature Microbiology*, 3(7):804–813, July 2018.

[4] Samuel Chaffron, Hubert Rehrauer, Jakob Pernthaler, and Christian von Mering. A global network of coexisting microbes from environmental and whole-genome sequence data. *Genome Research*, 20(7):947–959, July 2010. 00242.

[5] Tal Korem, David Zeevi, Jotham Suez, Adina Weinberger, Tali Avnit-Sagi, et al. Growth dynamics of gut microbiota in health and disease inferred from single metagenomic samples. *Science (New York, N.Y.)*, 349(6252):1101–1106, September 2015.

[6] Akintunde Emiola and Julia Oh. High throughput in situ metagenomic measurement of bacterial replication at ultra-low sequencing coverage. *Nature Communications*, 9(1):4956, November 2018.

[7] Nacho Molina and Erik van Nimwegen. Scaling laws in functional genome content across prokaryotic clades and lifestyles. *Trends in Genetics*, 25(6):243–247, June 2009.

Comment annoter et analyser les protéines à motifs répétés : Cas des protéines contenant des répétitions riches en leucine (LRR) chez le riz.

Céline GOTTIN[1], Anne DIEVART[2], Nathalie CHANTRET[3] et Vincent RANWEZ[1]

[1] Montpellier Supagro, UMR AGAP, Montpelleir, France
[2] CIRAD, UMR AGAP, Montpellier, France
[3] INRA, UMR AGAP, Montpellier, France

Corresponding Author: Céline Gottin (celine.gottin@supagro.fr)

L'annotation et l'analyse des protéines contenant des motifs répétés est un challenge en bio-informatique. Le caractère répété de ces motifs conduit souvent à des alignements de séquences multiples de mauvaises qualité limitant les tentatives d'analyses lorsque les séquences deviennent divergentes. Une méthode pour éviter cette difficulté consiste à annoter chacune des répétitions des protéines afin de guider les alignements. Cette annotation passe généralement par une recherche de motifs grâce à des profils HMM (Hidden Markov Model) issus de bases de données (Pfam, SMART). Cependant, certains motifs répétés restent difficiles à annoter, limitant les analyses de certaines familles de gènes. C'est le cas des protéines contenant des répétitions LRR (Leucine-Rich Repeat).

Trois grandes familles de gènes sont impliquées dans l'immunité chez les plantes grâce à leur capacité d'interaction protéine-ligand : Les LRR-RLP, les LRR-RLK et les NBS-LRR [1]. La spécificité de ces protéines vis-à-vis d'un ligand est portée par le domaine LRR qui consiste en une répétition en tandem de motifs LRR (de 4 à plus de 30 unités). Un motif LRR est constitué de 20 à 30 acides aminés et est caractérisé par une partie très conservée de 11 ou 12 acides aminés (LxxLxLxxNxL ou LxxLxLxxCxxL) suivi d'une partie plus variable [2]. De nombreuses publications se sont intéressées à l'émergence et l'évolution de ces différentes familles de gènes au sein de différentes espèces. Cependant, limitées par la qualité des alignements des séquences répétées, l'ensemble de ces études ne s'intéresse qu'à une famille de gènes à la fois en se basant sur les alignements des domaines associés aux LRR : le domaine kinase pour les LRR-RLK et le domaine NB-ARC pour les NLR [3,4]. Ainsi, il existe peu de connaissances sur l'évolution des domaines LRR de ces 3 familles, sur les LRR-RLP (à cause de l'absence de domaine associé aux LRR) ou sur la variabilité des motifs LRR au sein d'une espèce.

Afin de pouvoir mener une étude évolutive plus globale des motifs LRR et de ces protéines chez le riz, nous avons développé une méthode de recherche plus exhaustive et automatique des motifs LRR, utilisable à l'échelle d'un protéome entier. Cette méthode consiste à améliorer la spécificité d'un profil HMM LRR vis-à-vis d'une espèce d'intérêt par un processus de recherche itératif. Testée sur des protéomes d'*Arabidopsis thaliana* et *Oryza sativa*, cette méthode a permis de mieux détecter les protéines contenant des LRR mais aussi d'identifier un maximum de motifs au sein de chaque protéine. L'extraction des motifs LRR d'un protéome nous permettra dans un premier temps d'étudier la variabilité de ces motifs au sein d'une espèce. Ces connaissances seront ensuite utilisables afin d'étudier l'évolution de l'ensemble des protéines LRR en développant une méthode d'alignement adaptée à la structure de ces domaines.

References

1. M K Sekhwal, P Li, I Lam, X Wang, S Cloutier and F M You. Disease Resistance Gene Analogs (RGAs) in Plants. *International Journal of Molecular Sciences*, (16/8):19248–19290, 2015.
2. A V Kajava. Structural diversity of leucine-rich repeat proteins. *Journal of Molecular Biology*, (277/3):519–527, 1998.
3. J-F Dufayard, M Bettembourg, I Fisher, G Droc, E Guiderdoni, C Perin, N Chantret and A Dievart. New Insight on Leucine-Rich Repeats Receptor-Like Kinase Orthologous Relationships in Angiosperms. *Frontiers in Plant Science*, (8/381), 2017
4. Z-Q Shao, J-Y Xue, P Wu, Y-M Zhang, Y Wu, Y-Y Hang, B Wang and J-Q Chen. Large-Scale Analyses of Angiosperm Nucleotide-Binding-Site-Leucine-Rich Repeat Genes Reveal Three Anciently Diverged Classes with Distinct Evolutionary Patterns. *Plant Physiology*, (170/4):2095–2109, 2016.

Comment prédire un gRNA efficace dans des contextes expérimentaux variés ? En apprenant des gRNA publiés

Ségolène Diry[1] and Virginie Chesnais*[1]

[1]LifeSoft – Entreprise privée – France

Résumé

Keywords CRISPR, gRNA, méta-analyse

Introduction

Le système CRISPR/Cas9 est une technologie d'édition de génome, permettant de modifier l'ADN d'une cellule grâce à l'utilisation d'une courte séquence d'ARN guide (ou gRNA) qui va servir de point d'ancrage à une endonucléase (protéine Cas) pour couper l'ADN. Il existe plusieurs types de protéines Cas nécessitant chacune un gRNA répondant à son propre pattern. Parmi les plus utilisées, on retrouve la protéine Cas9sp qui nécessite l'utilisation d'un gRNA de 20 nucléotides suivis par un motif PAM NGG. Pour cette Cas, il existe sur le génome humain 303.669.088 séquence répondant à ce pattern soit 1 gRNA toutes les 10 pb. Il a été montré que toutes les séquences dans une région donnée, à l'échelle d'un gène par exemple, n'avaient pas la même efficacité. Il est donc nécessaire de déterminer des critères permettant de de choisir le gRNA ayant la meilleure probabilité de couper l'ADN.

Ainsi de nombreux algorithmes, publiés ces dernières années, tentent de prédire l'efficacité des expériences CRISPR en étudiant les caractéristiques des gRNAs ayant le mieux fonctionné dans un modèle expérimental (1–5). Néanmoins, ces études portent le plus souvent sur des jeux de données assez restreints. Par exemple, les scores de Doench (1,2) ont été développés à partir de gRNAs dont l'efficacité a été évaluée sur 2 lignées cellulaires humaines porteuses de mutations, tandis que le score CRISPR SCAN (4) repose sur l'évaluation de gRNA dans des embryons de poisson zèbre. De même, la mesure de l'efficacité du système est parfois réalisée grâce à des systèmes rapporteurs indirects ne reflétant pas l'efficacité de coupure du système dans un contexte génomique réel (3). Nous nous sommes donc demandé : i. quelle était la pertinence de ces modèles face à des gRNA utilisés dans des contextes expérimentaux différents ; ii. si le modèle expérimental avait un impact direct sur les caractéristiques permettant de prédire l'efficacité d'un gRNA.

Pour répondre à cette problématique, nous avons généré un jeu de données par le biais d'une méta-analyse bibliographique, dans le but de récupérer des gRNAs utilisés dans des modèles expérimentaux variés afin de refléter au mieux la diversité des utilisations.

Résultats

*Intervenant

Nous avons ainsi recensé 60 articles scientifiques publiés entre 2013 et 2018. Pour chaque publication nous avons récupéré les séquences des gRNA, de même que des informations concernant la méthodologie expérimentale utilisée (organisme, type cellulaire, méthode d'intégration de l'ARN guide et de la Cas9...) et l'efficacité rapportée dans le système utilisé (taux d'insertions et de délétion). Nous avons ainsi obtenu un jeu de données de 190 séquences d'ARN utilisés à la fois dans des modèles cellulaires (lignées ou cellules primaires) et des modèles *in vivo* ou *ex vivo* chez 6 espèces différentes. Pour chaque gRNA, en plus des annotations expérimentales issues des publications, nous avons calculé plusieurs métriques internes à la séquence du gRNA (ex. pourcentage de GC, entropies) ou concernant leur environnement génomique (ex. localisation dans une région codante) ainsi que les scores de prédictions de 5 algorithmes publiés dans la littérature entre 2014 et 2016 (1–5).

Afin de déterminer la performance des algorithmes de prédiction nous avons ensuite cherché à évaluer la corrélation entre ces scores de prédiction et l'efficacité réelle rapportée pour chacun des guides. Aucune corrélation n'a pu être observée entre l'efficacité rapportée par les utilisateurs et la valeur rapportée par ces scores de prédiction (pearson r 0.03 à 0.20). En restreignant l'analyse aux gRNAs utilisés exclusivement dans des modèles cellulaires humains, nous avons observé, comme attendu, une amélioration de la performance globale de certains scores tels que les score de Doench 2014 et 2016 mis au point sur des gRNAs évalués dans des lignées cellulaires humaines.

Nous avons ensuite comparé chacun des algorithmes de prédiction entre eux. De manière intéressante, bien que ces scores utilisent des variables souvent similaires (ex. pourcentage de GC de la séquence, composition nucléotidique à certaines positions, séquence du PAM), nous n'avons pu observer aucune corrélation majeure entre ces différents scores (pearson r 0.01 à 0.61). Nous avons alors regardé la capacité de 13 variables à prédire l'efficacité d'un gRNA parmi celles les plus couramment utilisées dans les scores de prédiction. Parmi elles, nous avons pu observer que 7/13 (54%) ne semblaient avoir aucun impact sur l'efficacité, 2/13 (15%) semblaient avoir un effet inverse de celui suggéré dans les modèles de prédiction et 4/13 (31%) semblaient avoir l'effet attendu sur l'efficacité. En regardant spécifiquement les gRNAs utilisés dans les modèles cellulaires humains, nous avons observé une amélioration des performances des variables avec 6/13 (47%) critères présentant la même tendance sur l'efficacité que celle supposée. Cette observation suggère l'importance des modèles expérimentaux pour la définition des variables permettant de prédire l'efficacité d'un gRNA.

Nous disposons dans notre jeu de données de 9 gRNAs utilisés dans 2 modèles cellulaires humains différents et avons pu observer une efficacité variable des gRNAs selon le modèle (différences d'efficacité comprise entre 12% et 75%). Nous avons ainsi pu confirmer l'importance du modèle cellulaire dans la prédiction de l'efficacité d'un gRNA. Nous avons alors évalué de nouveau les 13 variables prédictives précédentes en regard des informations sur le protocole expérimental. Nous avons ainsi pu observer que certains critères, bien que n'ayant aucun caractère prédictif de l'efficacité sur l'ensemble des gRNA, étaient relevant dans certaines conditions expérimentales. Par exemple, bien que la présence d'un G en position 1 n'impacte pas de manière globale l'efficacité, ce critère semble essentiel dans le cas d'une intégration par électroporation ou lipofection de plasmides codant pour la Cas9 et le gRNA mais pas dans le cas d'une intégration médiée par vecteur viral. Ces premières observations semblent confirmer l'importance de la prise en compte de l'approche expérimentale pour prédire l'efficacité d'un gRNA.

Conclusion

Ces travaux nous ont permis de mettre en évidence que l'efficacité d'un gRNA ne dépend pas seulement de critères internes à la séquence mais aussi du modèle expérimental. En effet, les efficacités variables de certains gRNAs en fonctions des conditions expérimentales, de même que le caractère prédictif de certaines variables dans des conditions spécifiques renforcent cette hypothèse. Nous avons également pu constater que les scores de prédictions publiés dans la littérature semblaient peu performants sur des données réelles probablement du fait

qu'ils ont été mis au point sur des jeux de données homogènes ne tenant pas compte de la diversité des modèles expérimentaux.

Au final, en se basant sur ces observations et en utilisant des algorithmes de machine learning, notre objectif sera de mettre au point un score de prédiction tenant compte à la fois du modèle expérimental et des paramètres liés à la séquence du gRNA et à son contexte génomique afin de prédire au mieux l'efficacité d'une expérience CRISPR.

References

1. Doench JG, Hartenian E, Graham DB, Tothova Z, Hegde M, Smith I, et al. Rational design of highly active sgRNAs for CRISPR-Cas9-mediated gene inactivation. Nat Biotechnol. 2014;32:1262–7.

2. Doench JG, Fusi N, Sullender M, Hegde M, Vaimberg EW, Donovan KF, et al. Optimized sgRNA design to maximize activity and minimize off-target effects of CRISPR-Cas9. Nat Biotechnol. 2016;34:184–91.

3. Chari R, Mali P, Moosburner M, Church GM. Unraveling CRISPR-Cas9 genome engineering parameters via a library-on-library approach. Nat Methods. 2015;12:823–6.

4. Moreno-Mateos MA, Vejnar CE, Beaudoin J-D, Fernandez JP, Mis EK, Khokha MK, et al. CRISPRscan: designing highly efficient sgRNAs for CRISPR-Cas9 targeting in vivo. Nat Methods. 2015;12:982–8.

5. Bae S, Kweon J, Kim HS, Kim J-S. Microhomology-based choice of Cas9 nuclease target sites. Nat Methods. 2014;11:705–6.

Comparaison des réseaux métaboliques de bactéries phytopathogènes

Ludovic Cottret, Caroline Baroukh, Stéphane Génin

LIPM, INRA-CNRS Occitanie-Toulouse

Corresponding Author: ludovic.cottret@inra.fr

Les bactéries appartenant au complexe d'espèces Ralstonia solanacearum sont responsables de la maladie du flétrissement chez de très nombreuses espèces de plantes. Elles sont regroupées en 4 phylotypes majeurs déterminés par la provenance géographique des souches (Asie, Amérique, Afrique, Asie du Sud Est + Australie). Au cours des dernières années, plusieurs souches appartenant aux 4 phylotypes ont été séquencées et leurs génomes comparés [2,3]. Par ailleurs, la reconstruction métabolique de la souche référence GMI1000 a permis d'analyser finement le lien entre métabolisme et virulence [1].

La question que nous nous posons actuellement est de savoir si des différences métaboliques existent entre les bactéries de ce complexe d'espèces et si ces différences peuvent être liées à leur mode d'infection, à leur spectre d'hôtes ou à leur phylotype.

Pour cela, grâce à un outil bioinformatique que nous avons développé, nous avons reconstruit les réseaux métaboliques de 19 espèces appartenant à ce complexe. Cette reconstruction s'est effectuée sur la base de liens d'orthologie avec 5 espèces dont le réseau métabolique a été expertisé manuellement. Comme ces réseaux de référence proviennent de sources différentes, une étape de standardisation des identifiants est nécessaire. Des mesures de qualité des réseaux basées sur leur topologie et la classification des réactions en voies métaboliques seront effectuées. Ensuite, la comparaison des réseaux métaboliques s'effectuera sur la base de leur contenu en voies métaboliques, en réactions et en métabolites en utilisant les méthodes statistiques d'analyses de correspondance et de clustering. Enfin, une interface interactive développée sous R grâce au package Shiny permettra de faciliter l'analyse de ces comparaisons par les biologistes. Plus tard, le raffinement de ces reconstructions permettra de comparer le fonctionnement de ces réseaux par analyse de graphes ou analyse de flux.

References

[1] Peyraud,R. et al. (2016) A Resource Allocation Trade-Off between Virulence and Proliferation Drives Metabolic Versatility in the Plant Pathogen Ralstonia solanacearum. PLoS Pathog., 12, e1005939.

[2] Remenant,B. et al. (2010) Genomes of three tomato pathogens within the Ralstonia solanacearum species complex reveal significant evolutionary divergence. BMC Genomics, 11, 379.

[3] Remenant,B. et al. (2011) Ralstonia syzygii, the Blood Disease Bacterium and some Asian R. solanacearum strains form a single genomic species despite divergent lifestyles. PLoS One, 6, e24356

Comparative genomics of *Rhizophagus irregularis*, *R. cerebriforme*, *R. diaphanus* and *Gigaspora rosea* highlights specific genetic features in Glomeromycotina.

Emmanuelle Morin[1], Shingo Miyauchi[1], Hélène San Clemente[2], Eric C.H. Chen[3], Alan Kuo[4], Igor V. Grigoriev[4], Bernard Henrissat[5], Christophe Roux[2], Nicolas Corradi[3] and Francis M. Martin[1]

[1] INRA, Unité Mixte de Recherche 1136 Interactions Arbres/Microorganismes, Laboratoire D'excellence ARBRE, Centre INRA-Grand Est-Nancy,54280, Champenoux, France

[2] Laboratoire de Recherche en Sciences Végétales, UPS, CNRS, 24 Chemin de Borde Rouge-Auzeville, Université de Toulouse Castanet-Tolosan 31326, Toulouse, France

[3] Department of Biology, University of Ottawa, ON K1N9A7, Ottawa, Canada

[4] US Department of Energy Joint Genome Institute, CA 94598, Walnut Creek, USA

[5] Architecture et Fonction des Macromolecules Biologiques, CNRS, Aix-Marseille Université, 13288, Marseille, France

Corresponding Author: emmanuelle.morin@inra.fr

- Glomeromycotina is a lineage of early diverging fungi that establish arbuscular mycorrhizal (AM) symbiosis with land plants. Despite their major ecological role, the genetic basis of their obligate mutualism remains largely unknown, hindering our understanding of their evolution and biology.
- We compared the genomes of Glomerales (*Rhizophagus irregularis*, *Rhizophagus diaphanus*, *Rhizophagus cerebriforme*) and Diversisporales (*Gigaspora rosea*) species, together with those of saprotrophic Mucoromycota, to identify gene families and processes associated with these lineages and to understand the molecular underpinning of their symbiotic lifestyle.
- Genomic features in Glomeromycotina appear to be very similar with a very high content in transposons and protein-coding genes, extensive duplications of protein kinase genes, and loss of genes coding for lignocellulose degradation, thiamin biosynthesis and cytosolic fatty acid synthase. Most symbiosis-related genes in *R. irregularis* and *G. rosea* are specific to Glomeromycotina. We also confirmed that the present species have a homokaryotic genome organisation.
- The high interspecific diversity of Glomeromycotina gene repertoires, affecting all known protein domains, as well as symbiosis-related orphan genes, may explain the known adaptation of Glomeromycotina to a wide range of environmental settings. Our findings contribute to an increasingly detailed portrait of genomic features defining the biology of AM fungi.

Acknowledgements

This work was supported by the Laboratory of Excellence ARBRE (ANR-11-LABX-0002-01)

References

1. Emmanuelle Morin, Shingo Miyauchi, Hélène San Clemente, Eric C.H. Chen *et al*. Comparative genomics of *Rhizophagus irregularis*, *R. cerebriforme*, *R. diaphanus* and *Gigaspora rosea* highlights specific genetic features in Glomeromycotina. *New Phytologist,* doi: 10.1111/nph.15687, 2019.
2. Eric CH Chen , Emmanuelle Morin *et al.* High intraspecific genome diversity in the model arbuscular mycorrhizal symbiont *Rhizophagus irregularis. New Phytologist,* doi: 10.1111/nph.14989, 2018.

Comparative microbial pangenomics to explore mobilome dynamics

Adelme Bazin[1], Guillaume Gautreau[1] , Claudine Medigue[1], Alexandra Calteau[1] and David Vallenet[1]

[1]LABGeM, Génomique Métabolique, CEA, Genoscope, Institut François Jacob, CNRS, Université d'Évry, Université Paris-Saclay, Evry, France.

Corresponding Author: abazin@genoscope.cns.fr

For the last decade, pangenomics has provided new tools for researchers to estimate genomic diversity by partitioning gene content in terms of *core* and *accessory* genome [1]. The *core* genome consists in ubiquitous genes within the taxonomic group being studied and the *accessory* genome are the genes present in one or some individuals but not all. However, the content of the *core* genome is highly dependent on the number of genomes included in a study limiting the relevance of comparisons between studies. Moreover, the concept of *accessory* genome lacks subtlety as it gathers genes with a large range of frequencies.

Recently, a new tool named PPanGGOLiN (Gautreau *et al.,* in preparation) [2] was developed to exploit gene neighborhood topology, gene frequency and population structure to classify gene families using a graph-based approach into a 3-class partitioning of *persistent*, which is a relaxed definition of the *core* genome, *shell* which corresponds to genes belonging to some individuals of the population and potentially associated to environmental adaptations, and *cloud* genome which are genes present at very low frequencies in the pangenome.

Using this new approach, we can compute the pangenome of any microbial clade with a sufficient number of genomes. Thus the variable regions of the pangenome which are defined as areas with mostly *non-persistent* genes in a genome can be detected. Using the graph structure provided by PPanGGOLiN, we can detect gene modules within each pangenome by integrating both co-occurrences in variable regions and co-localization information thus grouping *non-persistent* genes with a potential functional linkage [3].

Once modules are defined, we can study module conservation between species by searching for common groups of colocalized genes using subgraph mining methods on the different pangenome graphs. Then, we can explore mobilome dynamics in a comparative pangenomics approach by studying module conservation between clades.

The developed method can be applied to any clades with a sufficient number of genomes. This methodology applied on different collections of pangenomes may help to associate functional modules that are shared between individuals of the same taxonomic group and/or between individuals of different clades with common environmental factors or phenotypic traits.

References

[1] Tettelin, H., Masignani, V., Cieslewicz, M. J., Donati, C., Medini, D., Ward, N. L., ... & DeBoy, R. T. Genome analysis of multiple pathogenic isolates of Streptococcus agalactiae: implications for the microbial "pan-genome". *Proceedings of the National Academy of Sciences*, *102*(39)13950-13955. 2005.

[2] PPanGGOLiN: depicting microbial species diversity via a pangenome graph, https://github.com/ggautreau/PPanGGOLiN

[3] Overbeek, R., Fonstein, M., D'souza, M., Pusch, G. D., & Maltsev, N. The use of gene clusters to infer functional coupling. Proceedings of the National Academy of Sciences, 96(6), 2896-2901. 1999.

Comparison of efficiency of gene regulatory network inference algorithms on genomic and transcriptomic data

Lise Pomiès[1], Celine Brouard[1], Brigitte Mangin[2], Nicolas Langlade[2] and Simon de Givry[1]
[1] Unité de Mathématiques et Informatique Appliquées de Toulouse (MIAT) INRA : UR875, Chemin de Borde Rouge, 31320, Castanet Tolosan, France
[2] Laboratoire des interactions plantes micro-organismes (LIPM) INRA,CNRS : UMR2594, Chemin de Borde Rouge, 31320, Castanet Tolosan, France

Corresponding author: lise.pomies@inra.fr

One of the central targets of Systems Biology is to decipher the complex behavior of a living cell in its environment. A gene regulatory network is a simplified representation of the gene-level interactions. Network inference methods are powerful tools to understand such complex biological processes. A huge number of inference algorithms exist, and it could be difficult to identify an algorithm adapted to a specific experimental dataset.

In our case, we study the resistance of sunflower to drought, from transcriptomic and genomic data. We have the SNP measurements of 350 different sunflower genotypes and the expression measurements of 200 genes (mostly transcription factors) on these genotypes. To evaluate the behavior of different inference algorithms, we constructed 100 artificial datasets with biological properties close to the properties of our real dataset measured on sunflowers.

Generally, the quality of the results obtained with these algorithms on our artificial datasets was lower than the one obtained on the previous published datasets used to test each method. This could be explained by the simplicity of those test datasets compared to our artificial datasets which are closer to a real biological dataset. Those different inference methods provide a sorted list of gene-to-gene edges. It is interesting to highlight that globally the first 50 edges are composed of at least 60% of true orientated edges.

To go further in our analysis we decided to perform a meta-analysis of the results obtained with the different inference methods. Interestingly, to obtain better results with the meta-analysis than with the single best method, we must remove results obtained with methods having low quality results.

Acknowledgements

This work was supported by the "SUNRISE" project of the French National Research Agency (ANR-11-BTBR-0005, 2012-2019).

Comparison of variant callers for detection of large insertions in medical diagnosis context

Wesley DELAGE[1], Julien THEVENON[2] and Claire LEMAITRE[1]

[1] Genscale INRIA, Campus de Beaulieu, 263 Avenue Général Leclerc, 35042 Rennes, France
[2] UMR 5309 CNRS, Université Joseph Fourier - Institut Albert Bonniot, 38042 Grenoble,France

Corresponding Author: wesley.delage@inria.fr

1 Abstract

Medical diagnosis aims to identify and validate genetic variations that may be involved in genetic diseases. Depending on the origin and type of suspected structural variants (SV), different techniques are used: from short read sequencing for mutations below kb to aCGH for variations above 1kb [1]. Despite the birth in recent years of long reads sequencing, genome analysis in a medical context focuses on short reads sequencing of the patient mainly for financial reasons and sequencing error. The study of the exome in the medical setting is now one of the routines when a disease of genetic origin is suspected [2]. Exome sequencing, more affordable than genome sequencing, has been used to diagnose 20 to 40% of patients with rare Mendelian diseases [3]. SNPs and small indels are now detected with very high accuracy. However, despite the tools developed, confidence in the detection of structural variants greater than a few dozen base pairs is not sufficient to be integrated into the medical routine [4].

Since 2009, more than 70 tools for detecting structural variants have been developed based on read mapping to a reference genome (https://github.com/deaconjs/ThousandVariantCallersRepo/blob/master/SV.md), even the most recent tools developed based on alignment such as Svaba or Manta indicate a better ability to detect deletions than insertions [5,6]. Detecting insertion variants from short reads remains a challenge with two limiting factors that are the short size of the reads (100-125 bp) and the great diversity in size and structure of the insertion to be detected [7]. MindTheGap published in 2014 provides an alternative to variant calling by freeing itself from alignment to a reference genome, it allows the identification and assembly of insertions of any size [8]. In recent years, the tool has undergone many improvements, including the analysis of exome data for clinical diagnosis. We present here a comparison between MindTheGap3.0 and alignment-based variant callers for duplication and de novo insertion variants on whole exome and whole genome sequencing data from 50 to 1000 bp. The evaluation on simulated human exome and genome data shows that variant callers based on alignment have a recall and precision greater than 0.95 for small insertions. However, the recall and the precision for these tools critically drop to 0.10-0.20 once the insertion is larger than read size. MindTheGap3.0 performs the best variant calling on whole exome with a recall greater than 0.80 and a precision greater than 0.90. MindTheGap3.0 maintains the best performance with whole genome sequencing with a recall greater than 0.70 and a precision greater than 0.80. MindTheGap3.0 is not impacted by read length which allows it to find and assemble most of insertions regardless of the size, the genotypes (heterozygous and homozygous) and the insertion type (duplication or *de novo* insertion).

2 References

[1] Iafrate, A. J., Feuk, L., Rivera, M. N., Listewnik, M. L., Donahoe, P. K., Qi, Y., ... & Lee, C. (2004). Detection of large-scale variation in the human genome. *Nature genetics*, *36*(9), 949.

[2] Richards, S., Aziz, N., Bale, S., Bick, D., Das, S., Gastier-Foster, J., ... & Voelkerding, K. (2015). Standards and guidelines for the interpretation of sequence variants: a joint consensus recommendation of the American College of Medical Genetics and Genomics and the Association for Molecular Pathology. *Genetics in medicine*, *17*(5), 405.

[3] Lee,H.,Deignan, J. L., Dorrani, N., Strom, S. P., Kantarci, S., Quintero-Rivera, F., ... & Fox, M. (2014). Clinical exome sequencing for genetic identification of rare Mendelian disorders. *Jama*, *312*(18), 1880-1887.

[4] Roy, S., Coldren, C., Karunamurthy, A., Kip, N. S., Klee, E. W., Lincoln, S. E., ... & Wang, C. (2018). Standards and guidelines for validating next-generation sequencing bioinformatics pipelines: a joint recommendation of the Association for Molecular Pathology and the College of American Pathologists. *The Journal of Molecular Diagnostics*, *20*(1), 4-27.

[5] Wala, J. A., Bandopadhayay, P., Greenwald, N. F., O'Rourke, R., Sharpe, T., Stewart, C., ... & Nusbaum, C. (2018). SvABA: genome-wide detection of structural variants and indels by local assembly. *Genome research*, *28*(4), 581-591.

[6] Chen, X., Schulz-Trieglaff, O., Shaw, R., Barnes, B., Schlesinger, F., Källberg, M., ... & Saunders, C. T. (2015). Manta: rapid detection of structural variants and indels for germline and cancer sequencing applications. *Bioinformatics*, *32*(8), 1220-1222.

[7] Daber, R., Sukhadia, S., & Morrissette, J. J. (2013). Understanding the limitations of next generation sequencing informatics, an approach to clinical pipeline validation using artificial data sets. *Cancer genetics*, *206*(12), 441-448.

[8] Rizk, G., Gouin, A., Chikhi, R., & Lemaitre, C. (2014). MindTheGap: integrated detection and assembly of short and long insertions. *Bioinformatics*, *30*(24), 3451-3457.

Comparison of tolerogenic dendritic cells used in clinic with other *in vitro*-derived myeloid cells by epigenetic and transcriptomic analyses

Florian Berger[1,2], Eros Marin[1,2], Cynthia Fourgeux[1,2], Amandine Even[1,2], Laurence Bouchet-Delbos[1,2], Alice Mollé[1,2], Maria-Cristina Cuturi[1,2], Jeremie Poschmann[1,2,#] and Aurélie Moreau[1,2,#]

[1] Centre de Recherche en Transplantation et Immunologie UMR1064, INSERM, Université de Nantes, 30 Boulevard Jean Monnet, 44093, Nantes, France

[2] Unité de Transplantation Urologie Néphrologie (ITUN), 30 Boulevard Jean Monnet, 44093, Nantes, France

[#] Co-supervisors

Corresponding Author: aurelie.moreau@univ-nantes.fr

Cell therapy is a promising strategy to treat patients suffering from autoimmune or inflammatory diseases, or receiving a transplant [1]. We developed a protocol to generate human tolerogenic dendritic cell (DC), named ATDCs (Autologous Tolerogenic DCs), which are currently being tested in a first-in-man phase I/II clinical trial in patients undergoing kidney transplantation [2]. We recently discovered that ATDCs use a mechanism of suppression that radically distinguishes these cells from other cell-based immunotherapies under clinical investigation. The original feature of these human tolerogenic DCs is that they suppress T cell proliferation and expand regulatory T cells *in vitro* via their secretome and more specifically via their lactate production (manuscript in revision).

In this project, we aim to decipher the pathways and key molecules related to their mechanisms of action. For that, Chromatin Immunoprecipitation Sequencing (ChIP-seq) and RNA Sequencing (RNA-seq) were performed on ATDCs and other myeloid populations (control DCs, mature DCs, macrophages M1 and M2, and monocytes). Epigenetic and transcriptomic analyses of these different populations highlight the similarities and differences across these in vitro-derived myeloid populations.

This study will allow us to understand the mechanisms responsible for the tolerogenic activity of these cells and paves the way to the extension of their clinical application to other diseases. By comparing these results with public data sets from myeloid cells isolated from tumor microenvironment, it will also answer fundamental questions about the involvement of myeloid cells in the evasion of immune responses in the microenvironment of tumors.

References

1. Bluestone, Jeffrey A., Angus W. Thomson, Ethan M. Shevach, and Howard L. Weiner. “What Does the Future Hold for Cell-Based Tolerogenic Therapy?” Nature Reviews Immunology 7, no. 8 (August 2007): 650–54.

2. Marín, Eros, Maria Cristina Cuturi, and Aurélie Moreau. “Tolerogenic Dendritic Cells in Solid Organ Transplantation: Where Do We Stand?” Frontiers in Immunology 9 (2018).

Conciliation of process description and molecular interaction networks using logical properties of ontology

Vincent HENRY[1,2], Giulia BASSIGNANA[1,2], Violetta ZUJOVIC[2], Fabrizio DE VICO FALLANI[1,2], Olivier DAMERON[3], Ivan MOSZER[2] and Olivier COLLIOT[1,2]

[1] Inria, Aramis project-team, Paris, France

[2] ICM, Inserm U1127, CNRS UMR 7225, Sorbonne Université, F-75013, Paris, France

[3] Univ Rennes, CNRS, Inria, IRISA - UMR 6074, F-35000 Rennes, France

Corresponding Authors: vincent.henry@inria.fr - olivier.colliot@upmc.fr

Background: Systems biology is mostly based on network analysis. Biological networks can be represented in different ways, from molecular interaction networks (MIn; e.g. for genome regulation) to process description networks (PDn; e.g. for systems dynamics). The choice of the representation is a key element to provide consistent answers to an initial hypothesis. Usually, public resources (e.g. the STRING database, KEGG, Reactome…) provide a single network representation that is not necessarily appropriate to the expected analysis. Yet, all types of network representations are intrinsically structured and manageable by logical rules. Thus, it opens perspectives to switch between network representations.

Ontologies are able to manage knowledge and manipulate object properties using logical descriptions and rules. We hypothesize that they are a suitable framework to deal with these perspectives. Here we present an ontology-driven methodology that results in the addition of MIn properties to PDn.

Methods: We designed the Molecular Network Ontology (MNO), which contains 42 classes imported from the Systems Biology Ontology and the Biological interlocked Process Ontology for a) molecular reactions (e.g. binding, conversion or transcription) and b) molecular participants (e.g. gene, native gene product or converted gene product). Then, process classes were formally defined according to participant classes using the "has input", "has output", "positively mediated by" or "negatively mediated by" properties.

The macrophage signal transduction map (MSTM) is a curated PDn that contains 724 molecular reactions involving 1,353 participants. As a use case, the MSTM network was integrated into MNO: reactions and biochemical entities or genes from MSTN became individual instances of process classes and participant classes, respectively. Edges from MSTN were represented by MNO properties. Other information (identifiers, cross-references to the literature or databases) were kept as individual annotations. Then, logical rules were designed to infer molecular interactions from the initial process descriptions.

Finally, in order to validate the consistency of the logical reasoning results, we compared the MIn inferred by MNO and the MIn provided by STRING, after extraction of the set of genes contained in MSTM.

Results: MNO can fully integrate process description information into its classes, then logical rules can automatically enrich the initial PDn properties with consistent MIn properties.

Conclusion: MSTM manipulation by MNO showed that an ontology can integrate different molecular network representations from a single complex one. MNO does not work as a translator but adds new properties between molecular entities and keeps the initial ones. MNO is takes advantage of ontology abilities: the integration of knowledge as individual instances of formal classes and the enrichment of relationships thanks to logical reasoning. Ontologies can enrich but cannot create knowledge: MNO is thus able to infer molecular interaction properties from PDn, but is unable to infer process properties from MIn wherein processes are not described. Such a resource opens perspective to expand the choice of appropriate networks for systems biology analysis.

Conversion from quantitative model in SBML core to qualitative model in SBML qual

Athénaïs Vaginay[1], Malika Smail[2] and Taha Boukhobza[3]
[1] Université de Lorraine
[2] LORIA
[3] CRAN

Corresponding author: athenais.vaginay@loria.fr

A lot of different formalisms are used to model biological systems. They all have their own strengths and weaknesses. To encode these models, the SBML format is the *de facto* standard. This format is developed in a series of Levels that improve the format and fix problems as they occur. The current release is Level 3, which allows the definition of packages that extend the core format. One of these packages is named qual and is used to encode qualitative models. To the best of our knowledge, their is no automatic pipeline to convert a quantitative model (such as a set of differential equations) encoded in SBML core into a qualitative model encoded with the qual package. Here, we explore such a pipeline. It consists of solving numerically the differential equation system in order to retrieve the time course data of concentration of species on which we apply a discretisation. We then extract from these data a truth table that is used to find a fitting boolean model. We are taking as example a model of cell division of fission yeast, which has been studied both quantitatively (with a set of differential equations [1]) and qualitatively (with a boolean model [2]).

References

[1] Bela Novak, Zsuzsa Pataki, Andrea Ciliberto, and John J. Tyson. Mathematical model of the cell division cycle of fission yeast. 11(1):277.

[2] Maria I. Davidich and Stefan Bornholdt. Boolean network model predicts cell cycle sequence of fission yeast. 3(2):e1672.

CuteVariant

Un visualisateur de variants génétiques pour le diagnostic médical

Démo et Poster

Sacha SCHUTZ[1], Lucas BOURNEUF[2] and Pierre VIGNET[2]

[1] Laboratoire de génétique moléculaire et génomique, Rennes
[2] Université de Rennes 1

Corresponding Author: sacha@labsquare.org

Résumé

Cutevariant est un logiciel léger, libre et multiplateforme, dédié à la visualisation et l'interprétation des variants de séquençage haut débit.
Le logiciel prend en entrée différents formats de fichiers (par exemple des VCF annotés), qu'il intègre dans une base de données SQLite. Comme Variant-tools[1], Gemini[2] et SnpSift[3], Cutevariant dispose d'un DSL (Domain Specific Language) pour faire des requêtes sur les variants. Mais contrairement à eux ces requêtes peuvent être construites par l'intermédiaire de différents contrôleurs présents dans une interface graphique. Actuellement, il est possible de choisir les colonnes à afficher, de construire des filtres et d'appliquer des opérations ensemblistes sur les variants. Un système de *plug-in* est également disponible permettant aux utilisateurs de personnaliser l'interface en créant de nouveaux modules intégrés (par exemple l'inclusion de l'ontologie HPO). Enfin, contrairement aux solutions commerciales, le logiciel s'exécute localement sur toute machine équipée des systèmes d'exploitation GNU/Linux, Windows et macOS. L'interprétation des variants peut donc se faire localement sans avoir besoin de transférer de données médicales sensibles sur des serveurs tiers hors de contrôle.
Cutevariant est écrit en Python 3 et utilise le framework Qt5 pour son interface graphique (à l'aide du module Pyside2 développé dans le cadre du projet officiel *Qt for Python*). Les différents types de fichiers sont supportés grâce à des connecteurs, jouant le rôle d'interfaces entre les fichiers bruts et les formats d'entrée dédiés. Pour l'instant seulement les sorties de l'annotateur *snpEff* sont supportées.
Nous avons testé l'import d'un fichier vcf annoté avec snpEff provenant de l'échantillon Coriell NA128784. Ce vcf couvre tous les variants localisés dans les exons du chromosome 10. L'import a duré 1 min sur une machine classique, le même ordre de grandeur que variants tools et gemini. Le variant rs4244285 du gène *CYP2C19* connu dans cet échantillon pour modifier un site d'épissage a été identifié rapidement depuis l'interface.

Installation

Le code source est disponible sur github à l'adresse :

https://github.com/labsquare/cutevariant

Le logiciel est disponible sur pypi à l'adresse :

https://pypi.org/project/cutevariant/

References

[1] **Variant-tools** :F. Anthony et al. Integrated annotation and analysis of genetic variants from next-generation sequencing studies with variant tools *Bioinformatics 28 (3): 421-422, 2012*
[2] **Gemini** :Paila, Umadevi et al. GEMINI: Integrative Exploration of Genetic Variation and Genome Annotations *PLoS Computational Biology,* 2013
[3] **SnpSift** :Cingolani, P. et al. Using Drosophila melanogaster as a model for genotoxic chemical mutational studies with a new program, SnpSift *Frontiers Media SA,* 2012

Cypascan: an online tool for star allele calling in pharmacogenetics

Claire-Cécile BARROT[1,2], Jean-Baptiste WOILLARD[1,2] and Nicolas PICARD[1,2]

[1] INSERM UMR1248, Université de Limoges, Centre de Biologie et de Recherche en Santé, Rue du Pr Descottes, 87025, Limoges, France

[2] CHU de Limoges, Laboratoire de Pharmacologie, toxicologie et pharmacovigilance, 2 avenue Martin Luther King, 87042, Limoges, France

Corresponding Author: claire-cecile.barrot@unilim.fr

The "star" nomenclature is a standard to describe the different allelic versions of pharmacogenes. It was implemented in the 1990s by an international consortium and has been widely adopted in the field of pharmacogenetics, in particular for cytochromes P450 [1]. It is commonly in drug dosing guidelines [2-4]. Variant allele (*2, *3, etc.) corresponds to a gene version with one or more Short Nucleotide Variations (SNV) as compared to the reference (*1) allele sequence.

Currently, the most common way to determine the star allele of a given gene relies on targeted genotyping of tag SNVs. It is time-consuming and can only confirm or exclude the presence of specific alleles, leading to potential mistakes. In addition, while stars alleles definitions are available through public databases like PharmVar [5] or PharmGKB [6], transcripts-dependent positions rather than more suitable genomic positions are often used to refer to Tag SNV (*e.g.* 100C>T or P34S rather than chr22:42526694G>A). This is another cause of mistakes.

The French National Pharmacogenetics Network recently defined a list of genes of particular interest in PGx [7], including drug metabolizing enzymes, membrane transporters and pharmacodynamic targets. For each star allele of the panel, we computed GRCh37 and GRCh38 positions of related SNVs in a database. We then developed a PHP online software *Cypascan* (available at https://pharmaco.chu-limoges.fr/cypascan) dedicated to geneticists and pharmacologists. *Cypascan* allows users to select genes of interest among the panel. It then analyses combination of relevant SNVs from the NGS variant calling file (VCF) provided, and eventually provides comprehensive reports based on star allele nomenclature. *Cypascan* can also generate BED files of hotspots to be used in variant calling pipeline.

References

[1] Kalman LV, Black JL, Clinic M, Sw S, Bell GC. *Pharmacogenetic Allele Nomenclature: International Workgroup Recommendations for Test Result Reporting*. 2017

[2] Caudle K, Klein T, Hoffman J, Muller D, Whirl-Carrillo M, Gong L, et al. *Incorporation of Pharmacogenomics into Routine Clinical Practice: the Clinical Pharmacogenetics Implementation Consortium (CPIC) Guideline Development Process*. Curr Drug Metab [Internet]. 2014;15:209–17.

[3] Swen JJ, Nijenhuis M, De Boer A, Grandia L, Maitland-Van Der Zee AH, Mulder H, et al. *Pharmacogenetics: From bench to byte an update of guidelines*. Clin Pharmacol Ther [Internet]. Nature Publishing Group; 2011;89:662–73.

[4] Amstutz U, Carleton BC. *Pharmacogenetic testing: Time for clinical practice guidelines*. Clin Pharmacol Ther [Internet]. Nature Publishing Group; 2011;89:924–7.

[5] Pharmacogene Variation Consortium (PharmVar) at www.PharmVar.org

[6] Whirl-Carrillo M, McDonogh E, Herbet J, Gong L, Sangkuhl K, Thotn C, et al. *Pharmacogenomics Knowledge for Personlized Medicine*. Clin Pharmacol Therpeutics [Internet]. 2012;92:414–7.

[7] Picard N, Barin-Le Guellec C, Cunat S, Beaumais T, Evrard A, Fonrose X, et al. *A consensual panel for next-generation sequencing in pharmacogenetics: proposal from the French national network of pharmacogenetics (RNPGx)*; Annual congress of the French society of Pharmacology and Therapeutics (SFPT), June 12-14th, 2019 (Lyon). Abstract in: Fundamental and clinical pharmacology (in press)

De-centralized database: new challenges to design innovative contextualization algorithms

Axelle DURAND[1,2], Estelle GEFFARD[1,2], Rokhaya BA[1,2], Sophie LIMOU[1,2,3], Sophie BROUARD[1,2], Alexandre LOUPY[4,5], Nicolas VINCE[1,2] and Pierre-Antoine GOURRAUD[1,2]

[1] Centre de Recherche en Transplantation et Immunologie UMR 1064, INSERM, Université de Nantes, Nantes, France

[2] Institut de Transplantation Urologie Néphrologie (ITUN), CHU Nantes, Nantes, France

[3] Ecole Centrale de Nantes, Nantes, France

[4] Department of Nephrology-Transplantation Necker Hospital, Assistance Publique-Hôpitaux de Paris, University Paris Descartes, Paris, France

[5] Paris Cardiovascular Research Centre – Biostatistics Unit University Paris Descartes, UMR-S970, Paris, France

Corresponding Author: pierre-antoine.gourraud@univ-nantes.fr

Abstract *Data exchange in research projects involving multi-centers/multi-partners, led to a paradigm shift in data sharing system. Classically, centralized infrastructures are created for storing, processing or archiving information. In the GDPR era, these structures may no longer be suitable for collaborative health projects due to regulations on confidentiality of sensitive data. To meet the challenge of accessing and using these data while ensuring data security protection, we designed an on-site distributable database linked to a computation integrator where each center integrate this local module (database+integrator). Then, centers collect, store and control their own patients' data. The founding principle of the architecture is that no individual data circulates outside the centers. We apply this principle in the multi-centric KTD-INNOV (Kidney Transplantation Diagnostic INNOVation) and EU-TRAIN (EUr TRAnsp-INnov) projects. Both projects are designed to integrate large-scale systematic clinical and biological data of kidney transplanted patients in order to develop and validate a precision medicine application. This application is therefore connected to all databases without going through a centralized database, and only to collect summarized populational results. For example, if we consider BMI from a database (N=4824) which includes data from 2 different centers (N=2971 and N=1853); the mean (m=24.216) obtained from the centralized database is equal to the pondered mean (m=24.216) computed from the 2 local centers (m=24.189 and m=24.259). This illustrates that it is possible to decentralize data while preserving the same results. The new challenges are then to apply this approach to more complex calculation or algorithms. The distributed database solution consolidates data security and eases collaboration on multi-centric research projects, where each center can control and account for their own patients' data usage.*

Keywords De-centralized database, distributed database, data security, multi-centric research.

Deciphering the activation states of plasmacytoid dendritic cells, their dynamical relationships and their molecular regulation

Abdenour Abbas[1], Karima Naciri[1], Geoffray Brelurut[2], Nils Collinet[1], Denis Thieffry[2], Elena Tomasello[1], Marco Pettini[3], Marc Dalod[1] and Thien-Phong Vu Manh[1]

[1] Centre d'Immunologie de Marseille-Luminy (CIML) - Marseille, France,
[2] Institut de Biologie de l'Ecole Normale Supérieure (IBENS) - Paris, France,
[3] Centre de Physique Théorique (CPT) – Marseille, France

Corresponding Author: vumanh@ciml.univ-mrs.fr

Introduction

Functional heterogeneity exists within each type of immune cells. A remarkable example of this intra-type heterogeneity is the restriction of cytokine production to a minor fraction of activated cells, whatever the cytokine and cell types considered [1]. The mechanisms regulating this functional heterogeneity in each immune cell type largely remain to be identified. Recent technological innovations enable the characterization of transcriptomes at the single cell level. This approach revealed a heterogeneity already at ground state in certain immune cell types. Its biological meaning remains puzzling. Here, we aim at investigating the functional heterogeneity of immune cell types for cytokine production, by studying a striking example: type I interferon (IFN-I) production by plasmacytoid dendritic cells (pDC). More generally, we will extend this study by investigating how the tissue microenvironment of pDCs contributes to shape their activation states.

Objectives and approaches

The main question that we will address is what restricts IFN-I production to only a small fraction of splenic pDCs. One hypothesis could be that some pDCs are already poised for cytokine production at ground state, before encountering the virus.

To test this hypothesis, we will characterize the heterogeneity of splenic pDCs by combining high content flow and mass cytometry with single cell gene expression profiling at ground state or during infection with murine cytomegalovirus, and with advanced computational analyses and mathematical modeling. This will allow us to i) identify in an unbiased manner and deeply characterize at the molecular level splenic pDC activation states at ground state and during a viral infection, ii) infer the dynamical relationships between these pDC activation states by combining pseudotime computational analyses and kinetic experimental measurements, iii) identify the gene modules characteristic of each pDC activation state, as well as their associated biological processes and upstream regulators, iv) build a predictive mathematical model of pDC activation and test/refine this model based on iterative validation experiments.

Conclusion

Our interdisciplinary project integrates various types of approaches, including wet lab experimentations combining the use of mouse models, the generation of bulk and single cell transcriptomics (scRNAseq), epigenomics (scATACseq) and proteomics (CyTOF) data, bioinformatics analyses and mathematical modeling to decipher the activation states of plasmacytoid dendritic cells (pDCs), their relationships and their functional specialization, with the ultimate goal to advance our understanding of the roles of these cells in immune defenses against viral infections or cancer.

References

1. Zucchini N. et al. Individual plasmacytoid dendritic cells are major contributors to the production of multiple innate cytokines in an organ-specific manner during viral infection. *Int Immunol*. 2008 Jan;20(1):45-56.

Detection of transcriptional regulatory motifs specific to plant gene responses in stress conditions

Margot Correa1, Julien Rozière2, Cécile Guichard2, Marie-Laure Martin-Magniette2 and Véronique Brunaud2

[1] Laboratoire de Mathématiques et Modélisation d'Évry (LaMME), Université d'Évry Val d'Essonne, UMR CNRS 8071, ENSIIE, USC INRA, 23 boulevard de France, 91037, Évry Cedex, France

[2] Institute of Plant Sciences Paris Saclay (IPS2)-INRA: UMR1403, CNRS: UMR9213, Université Paris Sud-Paris XI, Université d'Évry-Val d'Essonne, Université Paris Saclay-bâtiment 630, rue Noetzlin, 91405, Orsay, France

Corresponding Author: margot.correa@math.cnrs.fr

Many studies try to understand molecular mechanisms of plant responses in stress conditions [1]. The aim of this work is to characterize genes involved in biotic and abiotic stress responses in order to identify specific genes regulated by one category of stress. Moreover to understand these gene regulations, we explore regulatory motifs detected to be specific to one stress category (abiotic or biotic).

For this approach we used (i) data from GEM2Net database [2], dedicated to the co-expression of Arabidopsis genes in stress response conditions and (ii) a homemade tool, PLMdetect [3], to find preferentially enriched motifs in promoters of Arabidopsis genes. In GEM2Net, 681 clusters of co-expressed genes were obtained after a clustering (based on mixture model) of near 400 stress conditions organized into 18 biotic and abiotic stress categories. Next, we used PLMdetect (PLM for Preferentially Located Motifs) on each cluster to identify candidate motifs as TFBS (Transcription Factor Binding Site) specific to stress responses. We also supported these results in crossing other experimental databases to identify transcription factors that bind these motifs.

Among the 11289 genes corresponding to the gathering of biotic stress clusters, we identified 8 PLMs into promoter regions of 989 genes. With the same approach, among the 12804 genes from abiotic stress clusters, we identify 16 PLMs into promoter regions of 5870 genes.
So, we identified 8 specific motifs to biotic stress and 16 specific motifs to abiotic stress as potential TFBS of each category. Each of these motif is characterized by a preferential position with a functional region into promoters. We used topGO to perform GO terms enrichment analysis of biotic and abiotic stress response genes compared to Arabidopsis genes. Our results revealed stress-related GO term enrichments in genes associated to these motifs. The biotic stress response genes are enriched in protein kinase activity, that are enzymes early implied in process of transduction signal in biotic stress responses. We also found that a majority of genes is present among target genes of transcriptional factors (experimental data from DAP-seq). This analysis shows the capacity of PLMdetect tool to detect transcriptional regulatory motifs for co-expressed genes and highlights the implication of these genes, associated to these motifs, in stress responses.

References

[1] Sylvain Jeandroz. Editorial: Plant Responses to Biotic and Abiotic Stresses: Lessons from Cell Signaling. *Frontiers in Plant Science* 8: 1772; 2017
[2] Rim Zaag. GEM2Net: from gene expression modeling to -omics networks, a new CATdb module to investigate Arabidopsis thaliana genes involved in stress response. *Nucleic Acids Research* 43 (Database issue): D1010-7, 2015.
[3] Virginie Bernard. Improved detection of motifs with preferential location in promoters. *Genome* (9): 739-752, 2010.

Detection of unknown genetically modified organisms (GMO) by statistical analysis of high-throughput sequencing data

Julie Hurel[1], Mathieu Roland[2], Sophie Schbath[3], Stéphanie Bougeard[1], Mauro Petrillo[4], Fabrice Touzain[1]

[1] *ANSES, Laboratoire de Ploufragan, 22440 Ploufragan, France*

[2] *ANSES, Laboratoire de la santé des végétaux, 49000 Angers, France*

[3] *INRA, Centre de Jouy-en-Josas, 78352 Jouy-en-Josas, France*

[4] *JRC, Joint Research Centre Ispra, 21027 Ispra, Italy*

Corresponding Author: julie.hurel@anses.fr

The European Community has adopted a very restrictive policy regarding the dissemination and use of Genetically Modified Organisms (GMOs), whose use in food is poorly accepted by consumers. Although a maximum threshold exists for a food to be labeled "GMO-free", they are easily detectable only by known GMOs. In recent years, not described GMOs have been produced whose sequence is unknown making them not detectable by current PCR approaches. To date, no detection method have been described for the detection of unknown GMOs. The method was developed using two sets of raw reads of the bacteria *Bacillus Subtilis*: the first related to a genetically modified genome and the second to a wild bacteria. First, we use a cleaning pipeline to sort the coding sequences (CDS) of the unknown GMO into two categories, the potential GMO inserts and the CDS of the wild genome. Then, two Blastn are performed on the pangenome of *Bacillus Subtilis* one on the CDS and one on the whole genome. The insert sequences of an unknown GMO have a different vocabulary from that of the wild genome from which it is derived. A genome has its own vocabulary, consisting of words. Each word is a set of nucleotides with predefined length such as "ATGCCT". We search over-represented or rare words in the wild genome to define specificities of its vocabulary and highlight the vocabulary of an insert that shows different word proportions. This difference is evaluated through a distance between the words vector of all CDS of the wild genome and the words vector of each candidate GMO insert CDS. We use Bray-Curtis distance after comparison with alternative distances and employ two types of calculation: one based on proportions of short words, the other based on frequencies of long words over-expressed in wild genome. Finally, a machine learning step is implemented to discriminate CDS of GMO inserts. The learning datasets correspond to the wild CDS found in sample and to a databank of known GMOs. This databank contains only filtered sequences that don't belong to the species of the wild genome. The sequences of candidates GMO inserts not matching pangenome after the two Blasts are used as prediction data. 12 machine learning methods were tested using the Caret package. The performance of these methods was compared to obtain the most efficient methods in terms of predicting sequences of GMO inserts (highest sensitivity and specificity). We retained Random Forest method for Machine Learning. It preserves most of the GMO inserts. Applied on a wild dataset, we got one false positive. Based on searched properties of the wild genome, we cannot exclude to find in results genes coming from recent horizontal gene transfer instead of GMO insert.

Development and validation of an alloscore in kidney transplantation

Amandine LECERF DEFER[1,2], Laurent MESNARD[3], Pierre-Antoine GOURRAUD[1,2], Nicolas VINCE[1,2] and Sophie LIMOU[1,2,4]

[1] ATIP-Avenir, Centre de Recherche en Transplantation et Immunologie UMR 1064, INSERM, Université de Nantes, Nantes, France

[2] Institut de Transplantation Urologie Néphrologie (ITUN), CHU Nantes, Nantes, France

[3] Unité mixte de recherche UMRS 1155, Inserm, Paris, France

[4] Ecole Centrale de Nantes, Nantes, France

Corresponding Author: nicolas.vince@univ-nantes.fr; Pierre-Antoine.Gourraud@univ-nantes.fr

Chronic kidney disease affects more than 10% of the population and is characterized by a progressive loss of kidney function that can lead gradually to end-stage kidney disease. Kidney transplantation is the best treatment against end-stage kidney disease and in 2015, 36,700 patients were living with a functional transplanted kidney in France. Short-term graft survival is well controlled by immunosuppressive treatments, but long-term survival remains insufficient (after 10 years, only 50% of grafts are still functional) and the molecular pathophysiological mechanisms leading to chronic rejection are still poorly understood. Donor-recipient HLA compatibility is the major factor associated with graft survival. Nevertheless, long-term graft failure, even for full-matched HLA pairs, suggests that additional factors beyond HLA could be immunogenic. In 2016, Mesnard and colleagues proposed a score, called allogenomic mismatch score [1], by summing non-synonymous amino acid differences on non-HLA transmembrane proteins between donor and recipient. This non-HLA alloscore obtained from whole-exome sequencin data (WES) focuses on rare variants and significantly correlates with the 3-yr post-transplantation renal function independently of HLA matching. Early this year, Reindl-Schwaighofer and colleagues developed a similar score (called SNP MM) from GWAS data focusing on common variants [2] and showed that the non-HLA SNP MM score significantly correlates with 5-yr graft survival.

In this context, we have implemented the alloscore in our own datasets (60 donor-recipient pairs in WES and imputed GWAS) in order to compare the different methods. We have assessed the weight of rare vs. common variants, imputed vs. non imputed variants, and HLA vs non-HLA factors on post-transplantation kidney function. From WES data, we imputed SNPs using the Haplotypes Reference Consortium (HRC) and created an imputed GWAS dataset. We computed the alloscore with Python and R scripts, and compared the different scores using R. Finally, we also evaluated the role played by donor-recipient genetic distance by implementing IBS (identity by state) using PLINK and principal component analyses using the Eigenstrat software.

Here, we replicated the previous findings and confirmed the importance of non-HLA mismatching for kidney graft function, independently from HLA matching. High non-HLA mismatch scores are correlated with a lower kidney function, a risk marker for chronic graft rejection. Our preliminary results still needs to be refined and confirmed in an independent cohort. Overall, these results can have a major impact on future donor selection as it may simultaneously improve graft allocation and reduce rejection risk.

References

1. Mesnard L, Muthukumar T, Burbach M, Li C, Shang H, Dadhania D, et al. (2016) Exome Sequencing and Prediction of Long-Term Kidney Allograft Function. *PLoS Comput Biol* 12(9): e1005088. https://doi.org/10.1371/journal.pcbi.1005088
2. Reindl-Schwaighofer R, Heinzel A, et al. (2019) Contribution of non-HLA incompatibility between donor and recipient to kidney allograft survival: genome-wide analysis in a prospective cohort. *Lancet* 2019; 393: 910–17. https://doi.org/10.1016/S0140-6736(18)32473-5

Development of a complete HLA analysis pipeline: HLA-Functional Immunogenomic eXploration (HLA-FIX)

Ayan IANNIELLO[1,2], Sophie LIMOU[1,2,3], Estelle GEFFARD[1,2], Venceslas DOUILLARD[1,2], Pierre-Antoine GOURRAUD[1,2] and Nicolas VINCE[1,2]

[1] ATIP-Avenir, Centre de Recherche en Transplantation et Immunologie UMR 1064, INSERM, Université de Nantes, Nantes, France

[2] Institut de Transplantation Urologie Néphrologie (ITUN), CHU Nantes, Nantes, France

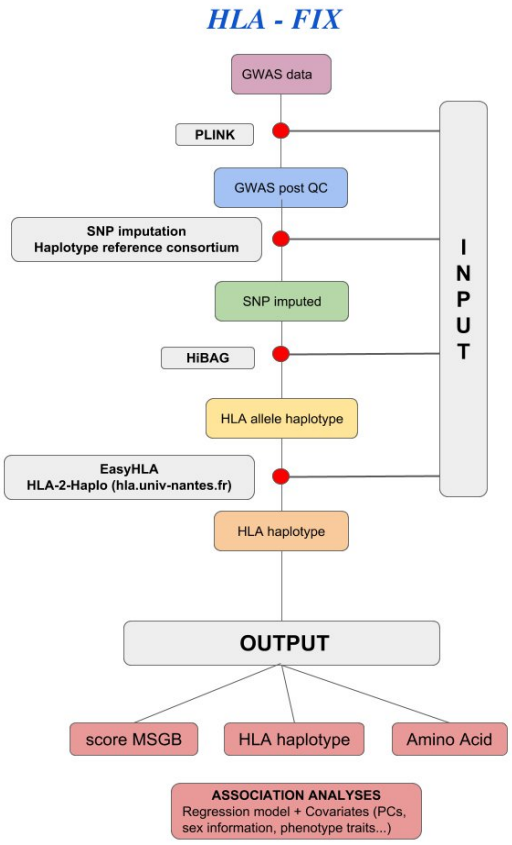
[3] Ecole Centrale de Nantes, Nantes, France

Corresponding Author: nicolas.vince@univ-nantes.fr

Genome wide association studies (GWAS) allowed a substantial increase of genetic association during the past years. These powerful analyses link traits to simple genetic markers called SNP (single nucleotide polymorphism) and give clues to understand molecular process involved in disease. SNPs in *HLA* region show association with more than 25% of GWAS catalog traits, mostly immune diseases and infectious phenotypes.

HLA region represents 1% of the genome but is the most polymorphic with thousands of alleles (21,499 alleles at this time according to IPD-IMGT/HLA Database). HLA genes code for proteins expressed on cell surface, they present short endogenous and exogenous peptide to T cells. In addition of body's immune response, HLA also plays a key role in the selection, differentiation and maturation of immune cells. These functions are just the visible part of the iceberg and its entire role is far from being fully understood. Defect or particular HLA haplotype can lead to unfavorable immune response such as in allergies, graft rejections, adverse drug reactions etc… Hence deep study of its role is crucial.

High resolution HLA-typing techniques remain expensive and time-consuming. A good alternative is statistic inference. In fact, by using SNP in linkage disequilibrium with *HLA*, we can efficiently infer *HLA* alleles. Many algorithms and pipelines exist to impute *HLA* alleles such as HiBAG R package. However analysis tool of HLA disease association are rare. R package BigDawg is one example of tool analysis but has a major drawback, it cannot include covariates. Here we developed a new HLA analysis pipeline on R, *HLA-FIX* to perform powerful *HLA* alleles association with traits (diseases or other phenotypes).



HLA-FIX can receive diverse input data: untreated GWAS data, post-imputation GWAS data, imputed *HLA* alleles or even typed *HLA* alleles. If needed *HLA-FIX* proceeds to all quality control and imputations steps using different tools high-cited in the literature. For that purpose it uses Plink for the quality control, then proceed to SNP imputation with Haplotype Reference Consortium imputation service, and impute *HLA* alleles with HiBAG. Then, *HLA-FIX* performs regression model analysis to test HLA alleles for association with given traits and can include covariates such as ancestry principal components or sex information. These potential confounding factors are essential in modern genetic analyses. Within a reasonable time, it returns the HLA allele, association p-value and size effect. Thus *HLA-FIX* will help to accelerate research and highlight interesting path for further exploration of immune related diseases.

Keywords

HLA, disease association, pipeline, imputation, -SNP, GWAS, immunology

Development of a novel multi-scale integrative computational method dedicated to the analysis of heterogeneous omics data.

Galadriel BRIERE[1], Ludovic LÉAUTÉ[1], Raluca URICARU[1] and Patricia THÉBAULT[1]

[1] Univ. Bordeaux, CNRS UMR 5800, LaBRI, France

Corresponding Author: thebault@labri.fr,uricaru@labri.fr

Data analysis, and not its production, has become the bottleneck in bioinformatics research. The integration of multiple types of omics data, by overlaying different points of view given by transcriptomics, proteomics and metabolomics analysis, provides a more insightful picture of life and allows to improve the quality of predictive models [1]. The common idea today is that the regulation of the cell occurs at numerous levels and therefore necessitates to carry out multi-scale analysis. Moreover, as a direct consequence of the present era marked by data massiveness, great support to achieve this is given by the large accessibility to a huge number of heterogeneous data sets [2]. To make profit of this large amount of data, becoming one of the biggest challenges in bioinformatics, many new computational methods [3] have been developed within the last two decades.

In this context, we developed Neomics, a hybrid model capable of representing a vast and rapidly evolving repository combining various types of information from biological (omics) data, in conjunction with the results of their computational analyses. Our model capitalizes on the methodological framework proposed by graph-oriented NoSQL databases, such as Neo4J, which will allow us to respond to the constraints of scalability and model dynamics. This conceptual framework is a solution of choice in the "data science" ecosystem but its adaptation to questions of science in bioinformatics is still marginal.

The richness of our model and of our visualisation system comes from the complete freedom the user has to adapt the model to his specific application. Moreover, the architecture of the graph database is not stuck in time; indeed the model is evolutive, thus allowing complete refactoring during the advancement of the project. Finally, the interactive interpretation of results, which is done by placing the expert in the center of the data analysis process, is a real added value. Guiding the analysis based on the effect of previous choices as well as on the interpretation of the data issued from the public databases, helps to improve the comprehension of the specificities and the integration of multiple types of omics data, and therefore encourages the generation of new hypotheses in the analysis process [4,5].

As an illustration, the biological questions that can be addressed with Neomics concern, among others, the identification of the main genes involved in a phenotype of interest (*e.g.* cancer), or the understanding of the mechanisms involved in the multi-scale level of the expression regulation of key molecules.

References

1. Palsson, *et al*. « The Challenges of Integrating Multi-Omic Data Sets ». 2010. Nature Chemical.
2. Gomez-Cabrero, *et al.* « Data integration in the era of omics: current and future challenges ». 2014. BMC Systems Biology 8.
3. Champion M, *et al.* Module Analysis Captures Pancancer Genetically and Epigenetically Deregulated Cancer Driver Genes for Smoking and Antiviral Response. EBioMedicine. 2018
4. Thébault P, *et al.* Advantages of mixing bioinformatics and visualization approaches for analyzing sRNA-mediated regulatory bacterial networks. Brief Bioinform. 2015.
5. Ayllón-Benítez A, *et al.* A new method for evaluating the impacts of semantic similarity measures on the annotation of gene sets. PLoS One. 2018.

keywords : single-cell, RNASeq, T cell CD8+, yellow fever

Divergent Clonal CD8+ T Cell Differentiation Establishes a Repertoire of Distinct Memory T Cell Clones Following Human Viral Infections

Jeff Mold[1*], Laurent Modolo[2*], Joanna Hrd[1+], Margherita Zamboni[1+], Anton Larsson[1], Patrik Sthl[1], Erik Borgstrm[3], Simone Picelli[1], Bjrn Reinius[14], Rickard Sandberg[1], Pedro Ru[1], Carlos Talavera-Lopez[1], Bjrn Andersson[1], Kim Blom[5], Johan Sandberg[5], Jakob Michaelsson[5], Franck Picard[2#] and Jonas Frisen[1#]

[1] Department of Cell and Molecular Biology, Karolinska Institute, SE-171 77 Stockholm, Sweden
[2] LBBE, UMR CNRS 5558, Universite Lyon 1, F-69622 Villeurbanne, France
[3] Science for Life Laboratory, Division of Gene Technology, KTH Royal Institute of Technology, SE-106 91 Stockholm, Sweden.
[4] Department of Medical Biochemistry and Biophysics, Karolinska Institute, 171 77 Stockholm, Sweden.
[5] Center for Infectious Medicine, Department of Medicine, Karolinska Institutet, Karolinska University Hospital Huddinge, 14186 Stockholm, Sweden
[6]

Corresponding author: laurent.modolo@ens-lyon.fr
∗ Equal contribution as first author
+ Equal contribution as second author # Equal contribution as last author

CD8+ T cells are essential for controlling viral and bacterial infections as well as malignant cell growth. The total CD8+ T cell response to a foreign antigen is composed of T cell clones with distinct T cell receptors generating a diverse array of functionally distinct cells. However, how individual clones combine to form this broad array is poorly understood, particularly in humans. We tracked CD8+ T cells by single cell RNAseq after a yellow fever vaccine shot. Using tools from ecological sciences and the scRNAseq literature, we performed a longitudinal clonal analysis. We demonstrated that clones exhibited biased differentiation towards effector or central memory T cell subsets, which gave rise to phenotypically distinct effector populations after secondary activation. This was confirmed in subsequent analysis of secondary responses to yellow fever or influenza infection. Our results demonstrate early clonal diversification and specialization of the CD8+ T cell responses to vaccines and after acute infections, identifying this as a hallmark of the CD8+ T cell response in humans.

Dynamic cell population modeling with UpPMaBoSS

Gautier Stoll[1,2], Aurélien Naldi[3], Vincent Noël[4], Éric Viara[5], Emmanuel Barillot[4], Guido Kroemer[1,2,6,7,8], Denis Thieffry[3] and Laurence Calzone[4]

[1] Equipe labellisée par la Ligue contre le cancer, Université Paris Descartes, Université Sorbonne Paris Cité, Université Paris Diderot, Sorbonne Université, INSERM U1138, Centre de Recherche des Cordeliers, Paris, France

[2] Metabolomics and Cell Biology Platforms, Gustave Roussy Cancer Campus Villejuif, France

[3] Institut de Biologie de l'ENS (IBENS), Département de biologie, École normale supérieure, CNRS, INSERM, Université PSL, 75005 Paris, France

[4] Institut Curie, Université PSL, INSERM, U900, F-75005, Paris, France

[5] Sysra

[6] Pôle de Biologie, Hôpital Europeen Georges Pompidou, AP-HP, Paris, France

[7] Suzhou Institute for Systems Biology, Chinese Academy of Sciences, Suzhou, China

[8] Department of Women's and Children's Health, Karolinska University Hospital, Stockholm, Sweden

Corresponding author: gautier.stoll@upmc.fr, laurence.calzone@curie.fr

1 Motivation

Mathematical modeling of biological perturbations (e.g., drug treatments) remains a challenge, especially when considering the dynamics of the cell population, taking into account events such as cell death, cell division, and cell-cell communication. Agent-based approaches are well suited when the spatial organization of cells is known, but inappropriate when it is poorly characterized.

2 Results

UpPMaBoSS is a new framework for dynamic cell population modeling. Relying on the preexisting tool MaBoSS, which enables probabilistic simulations of qualitative cellular networks, and adding a novel layer to account for cell interactions and population dynamics.

Here, we interpret the distribution of cellular state probabilities, estimated by MaBoSS, as the composition of an heterogeneous cell population. UpPMaBoSS alternates the simulation of individual cells with regular updates of the cell population and environmental signals. In particular, key network nodes are defined to account for cell division, cell death, and cell-cell interactions. During the population update phase, dead cells are removed from the population, while the probability of dividing cells is doubled. Finally, output components are integrated into a new probability distribution of input components for the next cellular update phase.

3 Illustration and discussion

We illustrate the use of UpPMaBoSS with a model of TNF-induced cell death, revealing a resistance mechanism. More generally, our probabilistic framework can be applied to many models of cellular networks, for example to study the impact of ligand release or drug treatments on cell fate decisions, such as commitment to proliferation, differentiation, apoptosis, etc. UpPMaBoSS simulations are relatively easy to encode and require only moderate computational power.

Développement et validation de pipelines pour l'analyse de données NGS dans le cadre du diagnostic en oncogénétique somatique

Nicolas Soirat[1,2,3], Anne-Laure Bouge[2], Jérôme Audoux[2], Sacha Beaumeunier[2], Charles Van Goethem[1], Julie Vendrell[1], Jérôme Solassol[1] et Nicolas Philippe[2]

[1] Laboratoire de Biologie des Tumeurs Solides, Hôpital Arnaud de Villeneuve, CHU de Montpellier, 371 Avenue du Doyen Gaston Giraud, 34000, Montpellier, France

[2] SeqOne, IRMB Hôpital Saint-Eloi 80 Avenue Augustin Fliche, 34090, Montpellier, France

[3] Parcours Bioinformatique, Connaissance, Données du Master Sciences Numérique pour la Santé, Université de Montpellier, Place Eugène Bataillon, 34095, Montpellier, France

Auteur référent: nicolas@soirat.fr

Abstract

Il est nécessaire d'utiliser des outils précis et efficaces afin de pouvoir détecter des variants parfois peu couverts, dans des zones à forte complexité, et ayant une faible fréquence allélique[1]. Dans l'étude des tumeurs solides, la conservation en paraffine rajoute un biais supplémentaire qui complexifie le rendu de diagnostic, en générant notamment des variants faux-positifs. Il est alors nécessaire d'appliquer des étapes supplémentaires lors du processus d'analyse et d'interprétation, d'utiliser des méthodologies différentes, d'élaborer un certain nombre de filtres[2] afin d'augmenter la sensibilité, la précision et la spécificité lors de la détection de variants pathogènes.

Avec l'avancée des technologies biologiques, le NGS devient un outil de choix en ce qui concerne le développement de nouvelles pistes thérapeutiques pour les tumeurs solides. Les thérapies ciblées deviennent une réalité, notamment grâce à des protocoles permettant d'inclure un large panel de régions, plus ou moins complexes (Alu, introniques...) et de différentes natures (génomique, transcriptomique, épigénomique...). Dans le cadre du Plan France Génomique, le défi est de pouvoir établir un diagnostic sur des génomes complets dans une routine clinique qui permet d'apporter des nouvelles pistes thérapeutiques, plus personnalisées, pour des patients qui sont en échec thérapeutique, dans une ère ou les essais cliniques se multiplient[3].

SeqOne propose des applications intégrant des briques d'outils académiques (i.e, MuTect2[4] et FreeBayes) avec une validation en routine clinique grâce à des contrôles positifs et des validations biologiques de laboratoire, en collaboration avec le CHU de Montpellier. Dans ce travail, nous nous intéressons plus particulièrement à de nouvelles approches WES (Whole Exome Sequencing) ou RNA-Seq pour identifier de nouvelles solutions pour des patients atteints d'un cancer.

Références

[1] Shany Koren and Mohamed Bentires-Alj. Breast Tumor Heterogeneity : Source of Fitness, Hurdle for Therapy. *Molecular Cell*, 60(4) :537–546, nov 2015.

[2] Samuel P Strom. Current practices and guidelines for clinical next-generation sequencing oncology testing. *Cancer biology & medicine*, 13(1) :3–11, mar 2016.

[3] David M. Hyman, Barry S. Taylor, and José Baselga. Implementing genome-driven oncology. *Cell*, (4) :584–599, feb.

[4] Scott L Carter, Matthew Meyerson, Michael S Lawrence, Gad Getz, Andrey Sivachenko, David Jaffe, Kristian Cibulskis, Stacey Gabriel, Eric S Lander, and Carrie Sougnez. Sensitive detection of somatic point mutations in impure and heterogeneous cancer samples. *Nature Biotechnology*, 31(3) :213–219, mar 2013.

Easy16S : a user-friendly Shiny interface for analysis and visualization of metagenomic data

Cédric MIDOUX[1, 2], Mahendra MARIADASSOU[2], Olivier RUE[2], Olivier CHAPLEUR[1], Valentin LOUX[2], Ariane BIZE[1]

[1] HBAN, IRSTEA, 1 rue Pierre-Gilles de Gennes, CS 10030, 92761, Antony, France

[2] MaIAGE, INRA, Université Paris-Saclay, Domaine de Vilvert, 78350, Jouy-en-Josas, France

Corresponding Author: cedric.midoux@irstea.fr

Microbiome data investigation has become a crucial step of recent studies of microbial diversity and dynamics. Studying microbial communities through NGS henceforth often involves the analysis and interpretation of large and high-dimensional datasets. For metabarcoding approaches, a two-step process is usually implemented. Bioinformatics processing of nucleotide sequence files is firstly performed to obtain, after several operations, count and affiliation tables. Secondly, statistical analyses and visualizations are classically used to explore the data and support interpretation. Such marker-gene sequencing approaches are currently affordable for most laboratories. They are used well beyond the community of bioinformaticians. Therefore, there is presently a high demand for user-friendly, interactive tools favoring the accessibility of data analysis to researchers with biology background.

Regarding the bioinformatics aspects, several solutions are available with a command-line approach or through the Galaxy platform (*ie*: FROGS [1]). We developed a tool for the second step: statistical analysis and visualization. To facilitate a quick and dynamic visualization of such data, we developed an interactive R-Shiny interface [2] named “Easy16S”. This tool is intended for biologists eager to explore their data and create figures rapidly and interactively. It is simple, easy-to-use and specifically focused on the mapping of covariates of interest.

To use Easy16S, an abundance data file in the biom format is required as primary input. Metadata (in tabular format) as well as a phylogenetic tree (nwk format) can also be added. After import, the data are available as a downloadable phyloseq object [3]. It is subsequently possible to plot various figures. Some statistical add-ons are also present. Currently, Easy16S supports count summaries, taxonomic and sample tables; community histograms; heatmap visualization of the count table; rarefaction curves; α-diversity and β-diversity plots; various multivariate analyses; phylogenetic tree browser and hierarchical clustering of communities.

Easy16S is mainly based on two R packages, shinydashboard [2] and phyloseq [3]. As it avoids the use of R command lines, it provides access to state-of-the-art methods and tools in the field and gives access to the R code which was executed at each step. All the figures can be adjusted with imported metadata. For example, covariates of interest can be mapped to color or shape, and samples can be grouped according a covariate of interest. Plots and tables can be exported in both raster and vector formats.

This application enabled biologists to explore data, to plot figures with specific covariates highlighted and to perform statistical analyses, in a simple and interactive way. It has already been used for a user-friendly integration and visualization of metabarcoding data.

Easy16S is currently run on an Open Source Shiny Server installed on the INRA MIGALE bioinformatics platform (http://genome.jouy.inra.fr/shiny/easy16S/). This project is currently managed in an IRSTEA GitLab repository (https://gitlab.irstea.fr/cedric.midoux/easy16S/). It was written with collaborative development and the continuous addition of features requested by users in mind. It is also open to suggestions from the community. The next steps for Easy16S project are a server resource optimization, authentication *via* a central user repository (LDAP) for data management and a full-fledged user manual.

1. Escudie, F., et al., *FROGS: Find, Rapidly, OTUs with Galaxy Solution.* Bioinformatics, 2018. **34**(8): p. 1287-1294.
2. Chang, W., et al., *shiny: Web Application Framework for R.* 2017.
3. McMurdie, P.J. and S. Holmes, *phyloseq: an R package for reproducible interactive analysis and graphics of microbiome census data.* PLoS One, 2013. **8**(4): p. e61217.

Eoulsan workflows for tag-based and full-transcript single-cell RNA-seq protocols

Geoffray BRELURUT[1, 2 *], Nathalie LEHMANN[1 *], Hatim EL JAZOULI[1, 3], Céline HERNANDEZ[1], Morgane THOMAS-CHOLLIER[1], Denis THIEFFRY[1], Stéphane LE CROM[1, 4], and Laurent JOURDREN[1]

[1] Institut de biologie de l'École normale supérieure, École normale supérieure, CNRS, Inserm, Université PSL, 75005 Paris, France

[2] Institut Curie, Université PSL, 75005 Paris, France

[3] Master Bioinformatique, Université de Rouen, 76130 Mont-Saint-Aignan, France

[4] Laboratory of Computational and Quantitative Biology, Institut de Biologie Paris-Seine, CNRS, Sorbonne Université, 75005 Paris, France

* These authors contributed equally to this work

Corresponding Authors: eljazoul@biologie.ens.fr & jourdren@biologie.ens.fr

Sequencing an unbiased snapshot of a given cell's transcriptome is now possible thanks to advances in handling minute amounts of genetic material and capturing individual cells. This progress enhanced the precision, sensitivity, and throughput of cellular expression profiling. Hence, these methods helped better characterize cell-to-cell heterogeneities and cell fates. Briefly, cell isolation and lysis, reverse transcription, and amplification are required before sequencing full-length transcripts (e.g., SMART-seq protocols) or one end of the transcript together with a unique molecular identifier (UMI) and a molecular barcode to, respectively, identify unique transcripts and cells (e.g., 10x Chromium/Drop-seq protocols). These two protocols produce data that require distinct preprocessing steps, the latter needing additional steps for UMI and barcode processing. The downstream analyses can however be performed likewise.

Implemented in Java, Eoulsan [1] is a modular workflow engine providing a reliable open-source framework for workflow management and reproduction, and therefore, offering an alternative to black-box commercial software and highly customized pipelines. The preprocessing steps take advantage of parallel computing by supporting popular job schedulers (Hadoop, TORQUE, or HTCondor). A Docker-Galaxy layout is used to ease the integration of new modules. The experimental design of a workflow is stored in a text file, while the workflow steps and parameters are listed in another XML file, ensuring flexibility and traceability. This approach allows to swiftly resume large analyses, and guarantees flexibility and reproducibility. A full documentation is available on GitHub [2].

The preprocessing steps include reads quality checking (FastQC), filtering reads, mapping reads to a genome (STAR, Bowtie), alignments quality checking, expression counting with reads or UMIs (HT-seq, featureCounts), and a MultiQC report. The results of all the aforementioned steps can be stored either in separate files, including a standard (dense or sparse) expression matrix, or in a Bioconductor object (SingleCellExperiment) containing the expression levels together with gene- and cell-specific metadata. Steps specific to tag-based protocols concern cell identification and filtering, and UMI processing (UMI-tools and soon, Alevin). For full-transcript protocols, the following cell-filtering methods are available: thresholding on raw metrics (e.g., number of detected features, read count), on median absolute deviation, and on sequencing saturation. Furthermore, read counts can be normalized using the deconvolution strategy (scran), or total count scaling. Results from downstream analyses such as differential gene expression (SCDE), cell clustering (Seurat), and cell fate reconstruction (Monocle 2) can be added to the generated object mentioned above. Complementary methods will be considered for implementation, depending on ongoing tests and community feedback.

In conclusion, the Eoulsan single-cell RNA-seq pipelines provide integrated workflows for analysing (on standalone workstations or on computer clusters) data stemming from the two main experimental protocols. The modular structure and the use of distributed data processing allow large amounts of data to be handled in a reproducible and flexible manner.

References

[1] https://outils.genomique.biologie.ens.fr/eoulsan/
[2] https://github.com/GenomicParisCentre/eoulsan/

Epigenome-wide association study reveals immunogenetic targets of DNA methylation modification by HIV-1

Abel Garnier[1,2], Nicolas Vince[1,2], George Nelson[3], Elizabeth Binns-Roemer[3], Victor David[3], Kevin Hoang[1], Pierre-Antoine Gourraud[1,2], James J. Goedert[4], Cheryl Winkler[3], Sophie Limou[1,2,5]

[1] ATIP-Avenir, Centre de Recherche en Transplantation et Immunologie UMR 1064, INSERM, Université de Nantes, Nantes, France

[2] Institut de Transplantation Urologie Néphrologie (ITUN), CHU Nantes, Nantes, France

[3] Frederick National Laboratory, Frederick, MD, United States of America

[4] NIH/NCI, Bethesda, MD, United States of America

[5] Ecole Centrale de Nantes, Nantes, France

Corresponding Author: Sophie.Limou@univ-nantes.fr

For nearly 25 years, extensive genetic and genomic association studies have revealed essential host factors for HIV control and disease progression, which notably led to the development of a new class of antiretroviral inhibitors (CCR5 antagonists). Overall, the identified associations account for ~20% of the phenotypic variance suggesting that other factors are yet to be discovered.

Here, we explore whether HIV-1 infection modifies the host epigenome DNA methylation patterns to identify host factors associated with HIV-infection. We recruited a unique collection of untreated HIV-infected individuals from the DC Gay cohort with longitudinal follow-up and PBMC samples available at pre-infection (n=23) and at several post-infection time points (n=57). Using the Illumina Infinium HumanMethylation450 arrays that cover over 485,000 methylation sites across the genome, we assessed the DNA methylation profiles of HIV infection adjusting for batch effect, age, cell composition, and population stratification.

Our analysis revealed that host genome DNA methylation profile is impacted by HIV-1 infection and highlighted several significantly differentially methylated sites ($P<10^{-7}$) which have been replicated in an independent cohort. These differentially methylated sites are located within genes previously known to exhibit an immune-related function or to interact with HIV-1 proteins, including the *HLA* and *PARP9*[1] genes. Our epignome-wide association study conducted in HIV-infected subjects has identified targets of epigenetic modifications by HIV-1, hence opening a new promising avenue for discovery of critical host factors interacting with the virus that might be leveraged for translation to drug or vaccine development.

References

1 König, R. *et al.* Global Analysis of Host-Pathogen Interactions that Regulate Early-Stage HIV-1 Replication. *Cell* **135,** 49–60 (2008).

Keywords

HIV, longitudinal cohort, EWAS, epigenetics, immunology, DNA methylation

Error Correction Schemes for DNA Storage with Nanopore Sequencing

Laura CONDE-CANENCIA[1], Belaid HAMOUM[1], Emeline ROUX[2,3] and Dominique LAVENIER[2]

[1] Lab-STICC, CNRS UMR 6285, Université Bretagne Sud, 56100, Lorient, France

[2] GenScale, INRIA, Campus Beaulieu, 35042, Rennes, France

[3] Calbinotox EA 7488, Université de Lorraine, Faculté des Sciences et Technologies Bd des Aiguillettes - B.P. 70239, F-54506 Vandoeuvre-lès-Nancy Cedex

Corresponding Author: laura.conde-canencia@univ-ubs.fr

1. Introduction

DNA storage is an emerging technology that uses DNA molecules to store data. This type of storage system is much more compact than any other due to the data density of the DNA. Moreover the capability for longevity and for resistance to obsolescence of DNA is undeniable: DNA is a universal and fundamental data storage mechanism in biology. For these and other reasons, DNA used as a memory-storage material in nucleic acid memory products promises a viable and compelling alternative to electronic memory. The exponential decrease in DNA synthesis costs should make the technology cost-effective for long-term data storage within about ten years. Several DNA-based storage systems have reported since 2012 [CHU12] [YON13] [GRA15] [ZHI16] [BOR16]. Companies such as Microsoft are leading research on this topic and have already announced their plan to use DNA storage in their data centers by 2020.

2. Motivation and goals of our work

The objective of our work is to design coding schemes allowing information to be efficiently encoded on DNA molecules, and to be read back using very low cost sequencing devices based on nanopore technology. The first step of our work develops coding techniques targeting the nanopore constraints in order to reduce the error performance of the global storage system. The second part demonstrates the feasibility of the approach by (1) synthetizing DNA molecules encoded with the proposed coding schemes; (2) reading the information by sequencing the DNA molecules with a nanopore device; (3) applying error detection and error correction techniques to the output signal to retrieve the initial information.

3. The MinION technology

We consider the Oxford Nanopore MinION device as it is currently the portable solution that offers ultra-long reads and a very reasonable cost. The MinION weighs under 100 g and can be plugged into a PC or a laptop using a high-speed USB 3.0 cable. It uses biomolecular nanopores from which electrical signals are detected when DNA molecules go through them. By interpreting these signals, DNA molecules can be deciphered. However, the nanopore technology presents high error rates (~10-15%), and so does the MinION. The challenge of our project is to provide efficient base-caller tools for correcting errors inherent to that technology. Our approach takes advantage of the DNA storage principle that allows for coding techniques to ensure robust decoding.

Acknowledgements

The authors would like to thank the Labex Cominlabs for funding this research project.

References

[CHU12] Church, G. M.; Gao, Y.; Kosuri, S. (2012), "Next-Generation Digital Information Storage in DNA" Science. 337 (6102): 1628.

[YON13] Yong, E. (2013). Synthetic double-helix faithfully stores Shakespeare's sonnets. Nature.

[GRA15] Grass, R. N.; Heckel, R.; Puddu, M.; Paunescu, D.; Stark, W. J. (2015). "Robust Chemical Preservation of Digital Information on DNA in Silica with Error-Correcting Codes". Angewandte Chemie International Edition. 54 (8): 2552.

[ZHI16] V. Zhirnov, R. M. Zadegan, G. S. Sandhu, G. M. Church and W. L. Hughes, "Nucleic acid memory", NATURE MATERIALS | VOL 15 | APRIL 2016 Online: http://arep.med.harvard.edu/pdf/Zhirnov_NAM_2016.pdf

[BOR16] James Bornholt, (2016). A DNA-Based Archival Storage System. ASPLOS '16 Proceedings of the Twenty-First International Conference on Architectural Support for Programming Languages and Operating Systems.

Etude de la composante génétique auto-immune de la Polyarthrite Rhumatoïde

Maëva VEYSSIERE[1], Fayrouz HAMMAL[1], Laetitia MICHOU[2], Jean-François DELEUZE[3], François CORNELIS[4], Elisabeth PETIT TEIXEIRA[1] and Valérie CHAUDRU[1]

[1] GenHotel, Univ Evry, University of Paris Saclay, Evry, France

[2]Division of Rheumatology, Department of Medicine, CHU de Québec-Université Laval - Québec, QC, Canada

[3]Centre National de Recherche en Génomique Humaine - François Jacob Institute, CEA - Evry, France

[4]GenHotel-Auvergne - Auvergne University, Genetic Department, CHU Clermont-Ferrand - Clermont-Ferrand, France

Corresponding Author: maeva.veyssiere@univ-evry.fr

1. Introduction

La Polyarthrite Rhumatoïde (PR) est un rhumatisme inflammatoire chronique qui touche environ 0,3% de la population française. Si le gène *HLA-DRB1*, facteur de risque génétique majeur, et une centaine de variants fréquents de type SNP ont été trouvés associés à cette maladie multifactorielle, ils n'expliquent pas la totalité de la composante génétique. Certaines recherches actuelles s'orientent ainsi vers l'identification de variants rares de type SNP, que les avancées dans les technologies de séquençage ont rendu possible. Un premier travail, réalisé à partir des données de séquences exome-entier de 30 individus appartenant à 9 familles à cas multiples de PR, nous a permis d'identifier un variant rare introduisant un codon-stop prématuré dans le gène SUPT20H [1]. Sachant que la composante auto-immune est importante dans le développement de la PR [2] et dans la mesure où d'autres maladies auto-immunes (MAI) sont présentes dans nos familles, notre objectif était d'identifier des variants rares pouvant être impliqués dans la composante génétique auto-immune et de caractériser des voies biologiques affectées par ces variants.

2. Matériel et méthodes

A partir des données de séquences de nos 30 individus (22 atteints de MAI dont la PR et 8 non atteints), nous avons réalisé des tests d'association-liaison avec l'outil pVAAST [3] qui permet de combiner le score d'association d'un *burden* test et les *lodscore* des variants d'un gène. Les gènes significativement associés (p-value < 0,05 après 1000 permutations) ont ensuite été filtrés pour ne garder que ceux pour lesquels au moins deux atteints de MAI différentes étaient porteurs d'un variant dans ce gène. Une analyse de sur-représentation a été réalisée à partir de cette liste finale avec CLUEGO [4].

3. Résultats

Sur les 819 gènes significativement associés à la composante auto-immune, 329 avaient un signal pouvant être attribué à plusieurs MAI. Six groupes fonctionnels, annotés dans KEGG, Reactome et Wikipathway, étaient sur-représentés dans cette liste de gènes ($p\text{-value}_{CLUEGO} < 0.03$). Enfin, pour 3 d'entre eux le signal d'association combiné, de l'ensemble des variants identifiés dans les gènes qui les composent, étaient significatif ($p\text{-value}_{pVAAST} \leq 0.03$) : mécanismes de réparation de l'ADN, O-glycosylation des protéines et Hépatite C. Notre analyse sera complétée par l'étude des intéractions Gène/Gène dans ces groupes fonctionnels.

Références

1. Veyssiere M, Perea J, Michou L, et al. A novel nonsense variant in SUPT20H gene associated with Rheumatoid Arthritis identified by Whole Exome Sequencing of multiplex families. PLOS ONE. 2019;14: e0213387.
2. Stanich JA, Carter JD, Whittum-Hudson J, Hudson AP. Rheumatoid arthritis: Disease or syndrome? Open Access Rheumatol Res Rev. 2009;1: 179–192.
3. Hu H, Roach JC, Coon H, et al. A unified test of linkage analysis and rare-variant association for analysis of pedigree sequence data. Nat Biotechnol. 2014;32: 663–669.
4. Bindea G, Mlecnik B, Hackl H, et al. ClueGO: a Cytoscape plug-in to decipher functionally grouped gene ontology and pathway annotation networks. Bioinformatics. 2009;25: 1091–1093.

Etude de la trajectoire de fréquences alléliques pahtogènes à travers le temps et l'espace

Marinna GAUDIN[1] and Christian DINA[1]

[1] l'institut du thorax, INSERM, CNRS, Univ Nantes, CHU Nantes, Nantes, France

Corresponding Author: marinna.gaudin@etu.univ-nantes.fr

L'étude de l'ADN ancien apporte un nouveau tournant à la compréhension de l'histoire évolutionnaire et démographique humaine. Depuis le XXe siècle, l'alliance de l'anthropologie et de la génétique constitue un outil puissant pour retracer les migrations du passé et pour identifier l'origine des populations contemporaines [1].

L'Europe fut la terre d'accueil de nombreuses vagues de migrations au cours des millénaires : les chasseurs-cueilleurs descendants des premiers Homo Sapiens (-15000), suivis des premiers fermiers depuis le Sud (-8500), et rejoints par des troupes nomades des steppes d'Europe de l'Est (-5000) [2]. Cependant, la population européenne contemporaine ne reflète pas un mélange homogène de ces anciennes communautés. Elle fait l'objet d'une *stratification génétique* prononcée à travers les régions [3], et ce à plusieurs niveaux d'échelle. La population française elle-même est finement stratifiée, si bien que l'on arrive à former des clusters génétiques qui subdivisent le territoire en 5 grandes régions : Nord-Ouest, Nord, Sud-Est, Sud-Ouest et Centre.

Nous possédons les données de génotypage d'individus du Mésolithique, Néolithique et Age de Bronze issus de trois études différentes [4,5,6], ainsi que les données d'individus modernes d'une étude française (SU.VI.MAX [7]) dont nous possédons les coordonnées géographiques. En comparant ADN ancien et ADN moderne, nous sommes en mesure d'étudier la trajectoire de fréquences alléliques de variants génétiques à travers le temps et l'espace. En un premier temps, nous chercherons à définir les proportions de populations sources ancestrales de ces différents clusters français. Puis, nous apporterons une analyse épidémiologique d'association des polymorphismes délétères selon le temps afin de suivre leur évolution, en lien avec des pathologies cardiovasculaires et le diabète de type II.

Cette étude nous permet de mesurer l'efficacité comparée de différentes méthodes d'estimation de partage génétique entre populations éloignées dans le temps : ACP, tests f3 et f4 ainsi que les modèles de mélanges.

Dans cette étude, nous démontrons que les populations du nord-ouest (Bretagne, Vendée) conservent des motifs génétiques relatifs aux chasseurs-cueilleurs, que les populations du nord montrent des liens de descendance plus forts avec les populations des steppes d'Europe de l'est, tandis que la structure génétique des régions du sud-ouest et du centre se réfère plutôt aux premiers fermiers.

References

1. « The Genomic Ancient DNA Revolution | Edge.org ». Consulté le 5 avril 2019. https://www.edge.org/conversation/david_reich-the-genomic-ancient-dna-revolution.
2. Mathieson et al. « Eight Thousand Years of Natural Selection in Europe ». *BioRxiv*, 10 octobre 2015. https://doi.org/10.1101/016477.
3. Sécher, Bernard. « Structure génétique à petite échelle de la population Française ». Text, 4 décembre 2015. http://secher.bernard.free.fr/blog/index.php?post/2015/08/27/Structure-g%C3%A9n%C3%A9tique-%C3%A0-petite-%C3%A9chelle-de-la-population-Fran%C3%A7aise.
4. Mathieson et al. « The Genomic History Of Southeastern Europe ». *BioRxiv*, 9 mai 2017, 135616. https://doi.org/10.1101/135616.
5. Lipson et al. « Parallel Palaeogenomic Transects Reveal Complex Genetic History of Early European Farmers ». *Nature* 551, n° 7680 (novembre 2017): 368-72. https://doi.org/10.1038/nature24476.
6. Olalde et al. « The Genomic History of the Iberian Peninsula over the Past 8000 Years ». *Science (New York, N.Y.)* 363, n° 6432 (15 2019): 1230-34. https://doi.org/10.1126/science.aav4040.
7. Hercberg, Serge et al. « The SU.VI.MAX Study: A Randomized, Placebo-Controlled Trial of the Health Effects of Antioxidant Vitamins and Minerals ». *Archives of Internal Medicine* 164, n° 21 (22 novembre 2004): 2335-42. https://doi.org/10.1001/archinte.164.21.2335.

Evolution of the angiotensin II receptors AT1 and AT2: Insights from molecular dynamics simulations

Asma Tiss*[1,2], Linda Grimaud[1], Rym Ben Boubaker[1], Laurent Marsollier[3], Hajer Guissouma[2], Daniel Henrion[4], and Marie Chabbert[4]

[1]Laboratoire MITOVASC – UMR CNRS 6015 - UMR INSERM 1083 , Université d'ANGERS – France
[2]Laboratoire LGIPH – Tunisie
[3]Centre de recherche en Cancérologie et Immunologie Nantes-Angers – INSERM U 1232 – France
[4]Laboratoire MITOVASC – UMR CNRS 6015 - UMR INSERM 1083 , Université d'ANGERS – France

Résumé

The renin-angiotensin system has a key role in cardiovascular and renal homeostasis. The octopeptide angiotensin II (Ang II) activates two G protein-coupled receptors, AT1 and AT2, sharing around 35% sequence identity. Most known effects of Ang II on the cardiovascular and renal systems are mediated by AT1. These effects include vasoconstriction, cardiac hypertrophy and contractibility and renal tubular sodium reuptake. AT2 is mainly expressed during development or in pathological conditions, and its effects might counterbalance excessive response of AT1. Due to the putative role of AT2 in cardioprotection and pain, AT2 may represent a valuable drug target. A better understanding of the structural differences between AT1 and AT2 is required to develop specific AT2 ligands. We carried out a phylogenetic analysis of the angiotensin receptors to highlight positions that were crucial for receptor evolution. This study revealed a S7.46N mutation in the sodium binding site of the AT1 receptor that occurred during the divergence of amniota. The sodium ion acts as a negative allosteric modulator of GPCRs. To investigate the effect of this mutation on the sodium binding properties of AT1 and its consequences on receptor functions, we carried out molecular dynamics simulations of the wild type AT1 and AT2 receptors and of the N7.46S AT1 mutant. This study revealed that, in any case, the sodium binding mode was dynamical within its binding cavity and moved between different sub-sites. Albeit the AT2 and N7.46S AT1 receptors possess the same residues in the sodium binding site, the sodium binding mode was different, indicating that residues from distant sites alter sodium binding properties. The implications of these results for the stability of the sodium binding site and the functions of the angiotensin II receptors are discussed.

*Intervenant

Exome sequencing in Hereditary Hypophosphatemic Rickets

Raphaël GAISNE[1,2,3], Lucile FIGUERES[2,3], Pascal HOUILLIER[4], Rosa VARGAS-POUSSOU [5], Sandrine LEMOINE[6], Claire LEMAN[1,2], Nicolas VINCE[1,2], Pierre-Antoine GOURRAUD[1,2], Sophie LIMOU[1,2,7]

1 ATIP-Avenir, Centre de Recherche en Transplantation et Immunologie UMR 1064, INSERM, Université de Nantes, Nantes, France
2 Institut de Transplantation Urologie Néphrologie (ITUN), CHU Nantes, Nantes, France
3 STEP, Regenerative Medicine and Skeleton UMRS 1229, Inserm, Nantes, France
4 Renal and metabolic Unit, Georges Pompidou Hospital, Paris, France
5 Departement of Genetics, Georges Pompidou Hospital, Paris, France
6 Nephrology department, Edouard Herriot Hospital, Lyon, France
7 Ecole Centrale de Nantes, Nantes, France

Corresponding Author: sophie.limou@univ-nantes.fr

Hereditary Hypophosphatemic Rickets (HHR) is a rare genetic disease causing nephrolithiasis and osteoporosis. Until now three genes (*SLC34A1*, *SLC34A3* and *SLC9A3R1*) have been identified but only account for a third of patients with HHR. In order to identify new candidate genes for HHR, we performed the first whole exome sequencing study on 26 unrelated HHR adult patients with no known genetic diagnosis, followed at the European Georges Pompidou Hospital (Paris), Edouard Herriot University Hospital (Lyon) and Nantes University Hospital.

Our 26 patients were predominantly males (69%), with familial history of nephrolithiasis (58%) and median age of onset of nephrolithiasis of 31 year-old. 11 patients (42%) had only nephrolithiasis, 3 had only osteoporosis (12%) and 12 patients (46%) were affected by both. Sequencing was performed at the Lille Integrated Genomic Advanced Network (CNRS platform) using an Illumina HiSeq 4000 with the Seq Cap MedExome capture kit (NimbleGen). We built our analysis pipeline with Snakemake from gold-standard tools. Reads were aligned with BWA MEM to hg19 and variants were called using GATK Haplotype Caller. The VCF file was then annotated with gnomAD, VEP, snpEff, mSigDB and ANNOVAR. Performances of the pipeline were assessed using the "NA12878" individual from the Genome In A Bottle consortium (recall 96.3%, precision 97.2% on "PASS" SNVs). The identified variants were filtered with SnpSift upon quality ("PASS", QUAL > 20), gnomAD allele frequency <1%, deleterious functional prediction (SnpEff, CADD, PROVEAN, FATHMM, REVEL) and linked with phosphate homeostasis according to mSigDB. No variant was shared by all patients nor was present in the three known genes of HHR. No variant was present exclusively in all patients sharing the same phenotype (nephrolithiasis, osteoporosis, both). In order to maximize the power for detecting novel variants causing HHR, we finally implemented a burden test using the gnomAD public database as previously described by Guo et al. [1]. At this stage, identification of a candidate variant by inheritance pattern is ongoing. Novel variants and candidate genes will be confirmed by Sanger sequencing.

In conclusion, we have implemented and validated a homemade pipeline dedicated to the analysis of whole-exome sequencing data. Burden test using gnomAD database was added to maximize the power for detection in the context of a limited case-only sample size. Our analysis has the potential to unravel novel genes causing the rare HHR disease and involved in renal tubular reabsorption of phosphate. Adding these genes to the current diagnostic panel will facilitate diagnosis confirmation and clinical management of HHR patients.

References

1. Guo MH, Plummer L, Chan Y-M, Hirschhorn JN, Lippincott MF. Burden Testing of Rare Variants Identified through Exome Sequencing via Publicly Available Control Data. The American Journal of Human Genetics. 2018 Oct;103(4):522–34.

Acknowledgments: *This work was supported by grant from the "Fondation du Rein"*

Keywords : *hereditary hypophosphatemic rickets, snakemake, whole exome sequencing, rare genetic disease*

Exploring relationship between to neuro-inflammatory diseases

Hadrien REGUE[1,2], Nicolas VINCE[1,2], Pierre-Antoine GOURRAUD[1,2] and David LAPLAUD[1,2]

[1] Centre de Recherche en Transplantation et Immunologie UMR 1064, INSERM, Université de Nantes, Nantes, France

[2] Institut de Transplantation Urologie Néphrologie (ITUN), CHU Nantes, Nantes, France

Corresponding Author: david.laplaud@univ-nantes.fr

Multiple sclerosis is a progressive auto-immune disease characterized by neuro-inflammation targeting the central nervous system. More than 100,000 people suffer from MS in France, with a majority of women (sex ratio close to 3:1). This disease is characterized by a strong ingress of immune cells from the blood to the cerebrospinal fluid and induce harsh handicaps [1]. Neuromyelitis Optica (NMO) is also a neuro-inflammatory disease which could lead to severe symptoms, close to the MS ones, if not treated carefully. As these two auto immune disease present similar clinical symptoms, the question of their genetic relationship was never asked previously.

In that purpose, we have collected blood from both MS and NMO patients, representing the largest NMO cohort assembled with more than 350 patients. All of these sample were genotyped using SNP microarray approach (Axiom PMRA chip, 800,000 SNPs). An update of the MS genetic map was published recently from the International Multiple Sclerosis Genetics Consortium (IMSGC), which reveal a total of 200 SNPs associated with MS [2]. From this list, we computed the multiple sclerosis genetic burden (MSGB) [3], a score which estimate the severity of MS based on the discovered SNPs within the MS genetic map. The computing of the MSGB score is based on an additive log model using the allelic Odds Ratio as a weight of each relevant SNPs, using the GWAS tool Plink and R programming language.

We measured the MSGB score in a subset of 95 individuals already genotyped. With limited power, we could replicate the MS genetic profile from Gourraud et al. We also calculated the MSGB score for NMO patients but could not observe the same trend, this may be due to lack of power. We plan to pursue our analyses with the full cohort. We will also test for association between MSGB score and different NMO phenotype subtypes.

Beyond this project, we are confident that our results will bring new clues about MS and NMO, but also establish a correlation between these two auto-immune diseases.

References

1. Salou, Marion, Bryan Nicol, Alexandra Garcia, et David-Axel Laplaud. « Involvement of CD8(+) T Cells in Multiple Sclerosis ». *Frontiers in Immunology* 6 (2015): 604. https://doi.org/10.3389/fimmu.2015.00604.
2. IMSGC. BioRxiv. 2017. http://biorxiv.org/lookup/doi/10.1101/143933

Gourraud, Pierre-Antoine, Joseph P. McElroy, Stacy J. Caillier, Britt A. Johnson, Adam Santaniello, Stephen L. Hauser, et Jorge R. Oksenberg. « Aggregation of Multiple Sclerosis Genetic Risk Variants in Multiple and Single Case Families ». *Annals of Neurology* 69, nº 1 (janvier 2011): 65-74. https://doi.org/10.1002/ana.22323.

Exploring white matter hyperintensities genetic associations through the use of external transcriptomic data

Elodie PERSYN[1], Matthew TRAYLOR[2], Hugh MARKUS[2] and Cathryn LEWIS[1,3]

[1] Department of Medical & Molecular Genetics, King's College London, Great Maze Pond, SE1 9RT, London, UK

[2] Department of Clinical Neurosciences, Stroke Research Group, University of Cambridge, R3 Box 83, Cambridge Biomedical Campus, CB2 0QQ, Cambridge, UK

[3] Social, Genetic and Developmental Psychiatry Centre, King's College London, de Crespigny Park, SE5 8AF, London, UK

Corresponding Author: elodie.persyn@kcl.ac.uk

From our large sample size genome-wide association study (GWAS) results on white matter hyperintensities (WMH), we aimed to integrate external transcriptomic data to identify and prioritize candidate genes.

White matter hyperintensities, a magnetic-resonance imaging (MRI) biomarker of small vessel disease, confer increased risk of stroke and dementia. The genetic study of WMH may provide insights into the underlying neurobiology, and identify much-needed potential treatment targets.

In order to discover new genetic associations with WMH, we performed a GWAS in European-ancestry UK Biobank samples on WMH (n=18,381). WMH GWAS results were meta-analysed with two multi-ethnic independent studies, from the CHARGE consortium (n=21,079) [1] and in stroke patients (n= 2,850) [2], for a total of 42,310 individuals. We identified 18 significantly associated loci for WMH, of which 9 loci were previously reported.

We further explored genetic association results by integrating external public transcriptomic data. Transcriptome-wide association studies (TWAS) [3], through the imputation of gene expression levels from GWAS results and external expression quantitative trait loci (eQTL) data, have been recently conducted to identify significant expression-trait association. We performed TWAS with FUSION [3] and identified 32 gene expression-trait associations for 6 selected vascular, blood and brain tissues. We performed colocalization analysis with COLOC [4] on these significant loci, as a complementary approach, to identify genes with colocalised GWAS and eQTL association signals. Among the 32 genes with significant expression-trait association, 21 genes present a high probability of sharing a GWAS and eQTL association signal.

These findings, through the integration of external functional data, will help prioritizing candidate genes for WMH and better identify molecular mechanisms underlying cerebral small vessel disease.

Acknowledgements

The authors thank the British Heart Foundation (RG/16/4/32218) for funding this study. We conducted our analyses by using the UK Biobank Resource (application no 36509) and the CHARGE GWAS summary statistics through the database of Genotypes and Phenotypes (dbGaP) (version: phs000930.v6.p1).

References

1. Benjamin FJ Verhaaren et al.. Multiethnic Genome-Wide Association Study of Cerebral White Matter Hyperintensities on MRI. *Circ Cardiovasc Genet.,* 8(2):398–409., 2015.
2. Matthew Traylor et al.. Genetic variation in PLEKHG1 is associated with white matter hyperintensities (n = 11,226). *Neurology*, 92(8):e749-e757, 2019.
3. Alexander Gusev et al.. Integrative approaches for large-scale transcriptome-wide association studies. *Nature Genetics*, 48(3):245–252, 2016.
4. Claudia Giambartolomei et al.. Bayesian Test for Colocalisation between Pairs of Genetic Association Studies Using Summary Statistics. *PLoS Genet.,* 10(5):e1004383, 2014.

JOBIM 2019 – Nantes

Poster

Title : Fast neutron variants detection in TILLING crop populations

Authors : Joseph Tran[1,2], Brahim Mania[1,2,3], Eulalie Lefeuvre[1,2], Fabien Marcel[1,2], Marion Dalmais[1,2], Abdelhafid Bendahmane[1,2]

1. Institute of Plant Sciences Paris Saclay IPS2, CNRS, INRA, Université Paris-Sud, Université Evry, Université Paris-Saclay, Bâtiment 630, 91405 Orsay, France.
2. Institute of Plant Sciences Paris Saclay IPS2, Paris Diderot, Sorbonne Paris-Cité, Bâtiment 630, 91405 Orsay, France.
3. Laboratoire de Génétique et Biométrie, Université ibn Tofaïl, B.P 242, Kénitra - Maroc

Abstract :

The EPITRANS platform in IPS2/Orsay has developped since 2008 a strong expertise in isolating targeted genes alleles in different crop species populations, a process known as TILLING. The platform uses EMS mutagenesis to produce such alleles (Dalmais et al. 2008; Boualem et al. 2008; Dalmais et al. 2013; Boualem et al. 2015; Roldan et al. 2017). To help screening a very large population (more than 2500 families per population), the platform has developped an efficient multi-dimensional screening method along with an analysis pipeline integrated in a desktop application called Sentinel (IDDN.FR001.240004.000.R.P.2016.000.10000). It fully automates the analysis workflow of short read sequencing libraries then saves the results to the backend database. It offers a user-friendly graphical interface to show the results in a meaningfull way, via tables and plots, to help biologists discriminate candidate point mutations (less than 1 % of the total results) from the background noise.
The platform has recently integrates a timely and complementary approach to generate diversity in crop populations using fast neutron mutagenesis. This kind of mutagenesis has been widely used to induce loss of function mutants in different plant model organisms (Rogers et al. 2009; Belfield et al. 2012) and crops (O'Rourke et al. 2013; Bolon et al. 2011; G. Li et al. 2017). But the screening methods were not efficient enough to identify large deletion and point mutations especially at the population scale. By using our very efficient screening method, allowing to identify all types of mutations induced in specific loci, we aim at identifying new fast neutron-induced alleles in large crop population. To achieve this goal, we built a new pipeline dedicated to detecting fast neutron-induced mutations, mostly indels as it was shown to be the most frequent type of mutation in rice (G. Li et al. 2017). It is now part of the new version of Sentinel.
Calling indels on mapped short paired-end reads to a reference sequence is much more challenging than SNP calling because of the indel itself which interferes with the mapping as most popular mapping approaches allow few missing base pairs (H. Li, Ruan, et Durbin 2008; H. Li et Durbin 2009). To reach a high sensitivity, we combined 4 indel calling methods : samtools, gatk, freebayes and pindel (H. Li et al. 2009; McKenna et al. 2010; Garrison et Marth 2012; Ye et al. 2009), at the cost of very high false positive rate. To overcome this problem, we combined the multi-dimensional screening with different quality filters to greatly reduce the false positive rate. We succesfully evaluated the pipeline performance on real and simulated datasets for small indels (≤ 30bp) and medium indels (> 30bp and ≤ 100bp). This suggests the pipeline is able to provide a good quality shortlist of candidate indels while limiting the false positive rate and the risk of missing true positives.
This pipeline is implemented using snakemake, an efficient workflow engine written in python, and conda, a package manager to keep track of all software dependencies in a dedicated research

environment. It is included in the latest version of Sentinel, a Lab Information Manager System desktop application written in C#.

keywords : fast neutron, mutation, indel, deletion, insertion, TILLING, crop, population, translational biology, reverse genetics, NGS, snakemake, conda, variant calling methods

EPITRANS platform website : http://ips2.u-psud.fr/fr/plateformes/epitrans-epigenomique-biologie-translationnelle.html

Definitions

TILLING (Targeted Induced Local Lesion IN Genome) is a molecular biology method developped to allow new mutations discovery in specific genes. It needs to combine a large scale of induced mutagenesis in a numerous population of interest, with a specifically sensitive DNA screening to seek rare mutations in a target gene.

References

Belfield, E. J., X. Gan, A. Mithani, C. Brown, C. Jiang, K. Franklin, E. Alvey, et al. 2012. « Genome-Wide Analysis of Mutations in Mutant Lineages Selected Following Fast-Neutron Irradiation Mutagenesis of Arabidopsis Thaliana ». *Genome Research* 22 (7): 1306-15. https://doi.org/10.1101/gr.131474.111.

Bolon, Yung-Tsi, William J. Haun, Wayne W. Xu, David Grant, Minviluz G. Stacey, Rex T. Nelson, Daniel J. Gerhardt, et al. 2011. « Phenotypic and Genomic Analyses of a Fast Neutron Mutant Population Resource in Soybean ». *Plant Physiology* 156 (1): 240-53. https://doi.org/10.1104/pp.110.170811.

Boualem, Adnane, Mohamed Fergany, Ronan Fernandez, Christelle Troadec, Antoine Martin, Halima Morin, Marie-Agnes Sari, et al. 2008. « A Conserved Mutation in an Ethylene Biosynthesis Enzyme Leads to Andromonoecy in Melons ». *Science (New York, N.Y.)* 321 (5890): 836-38. https://doi.org/10.1126/science.1159023.

Boualem, Adnane, Christelle Troadec, Céline Camps, Afef Lemhemdi, Halima Morin, Marie-Agnes Sari, Rina Fraenkel-Zagouri, et al. 2015. « A Cucurbit Androecy Gene Reveals How Unisexual Flowers Develop and Dioecy Emerges ». *Science (New York, N.Y.)* 350 (6261): 688-91. https://doi.org/10.1126/science.aac8370.

Dalmais, Marion, Sébastien Antelme, Séverine Ho-Yue-Kuang, Yin Wang, Olivier Darracq, Madeleine Bouvier d'Yvoire, Laurent Cézard, et al. 2013. « A TILLING Platform for Functional Genomics in Brachypodium Distachyon ». *PloS One* 8 (6): e65503. https://doi.org/10.1371/journal.pone.0065503.

Dalmais, Marion, Julien Schmidt, Christine Le Signor, Francoise Moussy, Judith Burstin, Vincent Savois, Gregoire Aubert, et al. 2008. « UTILLdb, a Pisum Sativum in Silico Forward and Reverse Genetics Tool ». *Genome Biology* 9 (2): R43. https://doi.org/10.1186/gb-2008-9-2-r43.

Garrison, Erik, et Gabor Marth. 2012. « Haplotype-based variant detection from short-read sequencing ». *arXiv:1207.3907 [q-bio]*, juillet. http://arxiv.org/abs/1207.3907.

Li, Guotian, Rashmi Jain, Mawsheng Chern, Nikki T. Pham, Joel A. Martin, Tong Wei, Wendy S. Schackwitz, et al. 2017. « The Sequences of 1504 Mutants in the Model Rice Variety Kitaake Facilitate Rapid Functional Genomic Studies ». *The Plant Cell* 29 (6): 1218-31. https://doi.org/10.1105/tpc.17.00154.

Li, Heng, et Richard Durbin. 2009. « Fast and Accurate Short Read Alignment with Burrows-Wheeler Transform ». *Bioinformatics (Oxford, England)* 25 (14): 1754-60. https://doi.org/10.1093/bioinformatics/btp324.

Li, Heng, Bob Handsaker, Alec Wysoker, Tim Fennell, Jue Ruan, Nils Homer, Gabor Marth, Goncalo Abecasis, Richard Durbin, et 1000 Genome Project Data Processing Subgroup. 2009. « The Sequence Alignment/Map Format and SAMtools ». *Bioinformatics (Oxford, England)* 25 (16): 2078-79. https://doi.org/10.1093/bioinformatics/btp352.

Li, Heng, Jue Ruan, et Richard Durbin. 2008. « Mapping Short DNA Sequencing Reads and Calling Variants Using Mapping Quality Scores ». *Genome Research* 18 (11): 1851-58. https://doi.org/10.1101/gr.078212.108.

McKenna, Aaron, Matthew Hanna, Eric Banks, Andrey Sivachenko, Kristian Cibulskis, Andrew Kernytsky, Kiran Garimella, et al. 2010. « The Genome Analysis Toolkit: A MapReduce Framework for Analyzing next-Generation DNA Sequencing Data ». *Genome Research* 20 (9): 1297-1303. https://doi.org/10.1101/gr.107524.110.

O'Rourke, Jamie A., Luis P. Iniguez, Bruna Bucciarelli, Jeffrey Roessler, Jeremy Schmutz, Phillip E. McClean, Scott A. Jackson, et al. 2013. « A Re-Sequencing Based Assessment of Genomic Heterogeneity and Fast Neutron-Induced Deletions in a Common Bean Cultivar ». *Frontiers in Plant Science* 4: 210. https://doi.org/10.3389/fpls.2013.00210.

Rogers, Christian, Jiangqi Wen, Rujin Chen, et Giles Oldroyd. 2009. « Deletion-Based Reverse Genetics in Medicago Truncatula ». *Plant Physiology* 151 (3): 1077-86. https://doi.org/10.1104/pp.109.142919.

Roldan, Maria Victoria Gomez, Claire Périlleux, Halima Morin, Samuel Huerga-Fernandez, David Latrasse, Moussa Benhamed, et Abdelhafid Bendahmane. 2017. « Natural and Induced Loss of Function Mutations in SlMBP21 MADS-Box Gene Led to Jointless-2 Phenotype in Tomato ». *Scientific Reports* 7 (1): 4402. https://doi.org/10.1038/s41598-017-04556-1.

Ye, Kai, Marcel H. Schulz, Quan Long, Rolf Apweiler, et Zemin Ning. 2009. « Pindel: A Pattern Growth Approach to Detect Break Points of Large Deletions and Medium Sized Insertions from Paired-End Short Reads ». *Bioinformatics (Oxford, England)* 25 (21): 2865-71. https://doi.org/10.1093/bioinformatics/btp394. 25 (21): 2865-71. https://doi.org/10.1093/bioinformatics/btp394.

Softwares

Bendahmane, A., Marcel F., Dalmais M., Beaumont G., Mania B. SENTINEL, SOFTWARE dedicated to TILLING by NGS Analysis. Certified by "Agence pour la Protection des programmes". Inter Deposit Digital Number.FR001.240004.000.R.P.2016.000.10000

Feedback on a comparative metatranscriptomic analysis

Cédric MIDOUX[1, 2], Tiago P. DELFORNO[3], Thais Z. MACEDO[4], Gileno V. LACERDA JR.[3], Olivier RUÉ[2], Mahendra MARIADASSOU[2], Maria B. A. VARESCHE[4], Théodore BOUCHEZ[1], Ariane BIZE[1], Valéria M. OLIVEIRA[3], Valentin LOUX[2]

[1] HBAN, IRSTEA, 1 rue Pierre-Gilles de Gennes, CS 10030, 92761, Antony, France
[2] MaIAGE, INRA, Université Paris-Saclay, Domaine de Vilvert, 78350, Jouy-en-Josas, France
[3] Microbial Resources Division, Research Center for Chemistry, Biology and Agriculture (CPQBA), Campinas University - UNICAMP, Campinas, SP CEP 13081-970, Brazil
[4] Laboratory of Biological Processes, Department of Hydraulics and Sanitation, Engineering School of São Carlos, University of São Paulo (EESC - USP) Campus II, São Carlos, SP CEP 13563-120, Brazil

Corresponding Author: cedric.midoux@irstea.fr

The progress of next generation sequencing favors the development of more comprehensive ecosystem studies thanks to metatranscriptomic approaches. These latter can indeed provide access to functional information at a good analysis depth. Through a study of anaerobic digesters treating anionic surfactant contaminated wastewater [1] (namely the linear alkylbenzene sulfonate, LAS), we developed a bioinformatics pipeline to perform the RNAseq data analysis for shotgun metatranscriptomics data.

In this pipe-line, the raw data are cleaned and pre-processed. Reads corresponding to rRNA are detected and discarded from the datasets. After a normalization step based on k-mer counts, the mRNA reads from the datasets are *de novo* co-assembled using the Trinity software. Coding regions of the metatranscriptomic assembly are subsequently predicted and annotated. For functional annotation, sequences with matches to the eggNOG and KEGG GENES databases are retrieved to establish functional categories and reconstruct the metabolic pathways. For taxonomic classification, the sequences are assigned by comparing them to a NCBI-nr database. For each dataset individually, reads are mapped back to the co-assembled contigs. Eventually, a count table is constructed; it contains, for each predicted gene, the counts obtained by samples, as well as the associated taxonomic and functional annotations.

After aggregation and statistical analysis, this study enabled detecting active genes likely involved in each step of LAS biodegradation and exploring the microbial active core related to LAS degradation.

1. Delforno, T.P., et al., *Comparative metatranscriptomic analysis of anaerobic digesters treating anionic surfactant contaminated wastewater*. Sci Total Environ, 2019. **649**: p. 482-494.

Flexible analysis of WGS of bacterial genomes using wgMLST approach.

Benoit VALOT[1], Charlotte COUCHOUD[1,2], Daniel MARTAK[1,2], Anaïs POTRON[1,3], Xavier BERTRAND[1,2], and Didier HOCQUET[1,2]

[1] UMR 6249 Chrono-Environnement, Université de Bourgogne Franche-Comté, 25000, Besançon, France

[2] Laboratoire d'Hygiène Hospitalière, Centre Hospitalier Régional Universitaire, 25000, Besançon, France

[3] Laboratoire de Bactériologie, Centre Hospitalier Régional Universitaire, 25000, Besançon, France

Corresponding Author: benoit.valot@univ-fcomte.fr

1 Background

Typing bacterial pathogens is an important public health task in hospital. The use of next generation sequencing to identify and follow epidimic clone is rising. For this purpose, core or whole genome Multilocus Sequence Typing (cgMLST or wgMLST, respectively) have become the new standard [1]. While the conventional MLST method relies on a few numbers (<10) of alleles located in housekeeping genes, cgMLST and wgMLST take into account the much larger the core or the whole genome of the collections. Common to all these methods, each unique sequence identifies a unique allele which combination determines the sequence type (ST) of the bacterial isolate.

2 Development

We developed pyMLST (python Mlst Local Search Tool) to perform this task automatically. Unlike existing tools, it uses a local sqlite database to store allele sequences and MLST profiles, allowing the expansion of the collection of compared genomes. The entry is either drafts genome produced by an assembler or genomes stored in sequence database.

The program is written in python on GPL3 license and source code is shared in github [2].

3 Applications

The performance of pyMLST were evaluated in three independent genome databases aiming at (*i*) deciphering *in vitro* evolution history of an *Escherichia coli* hypermutator strain [3], (*ii*) identifying independent outbreaks of *Pseudomonas aeruginosa* ST395, and (iii) characterizing a national outbreak of carbapenemase producing *Proteus mirabilis* in regards to the population structure of the whole species, retrieved from NCBI database.

References

[1] Maiden, Martin C. J., Melissa J. Jansen van Rensburg, James E. Bray, Sarah G. Earle, Suzanne A. Ford, Keith A. Jolley, et Noel D. McCarthy. *MLST revisited: the gene-by-gene approach to bacterial genomics*. Nature reviews. Microbiology, 2013.

[2] https://github.com/bvalot/pyMLST

[3] Ahrenfeldt, Johanne, Carina Skaarup, Henrik Hasman, Anders Gorm Pedersen, Frank Møller Aarestrup, et Ole Lund. *Bacterial whole genome-based phylogeny: construction of a new benchmarking dataset and assessment of some existing methods*. BMC Genomics. 2017.

Formatage et annotation des variants structuraux
-
Présentation du logiciel Svagga

Pierre-Antoine ROLLAT-FARNIER[1,2], Flavie DIGUET[2,3], Thomas SIMONET[1,4], Nicolas CHATRON[2,3], Damien SANLAVILLE[2,3], Claire BARDEL[1,5,6] and Caroline SCHLUTH-BOLARD[2,3]

1 Cellule Bioinformatique, Hospices Civils de Lyon, 69500, Bron, France
2 Service de Génétique, Hospices Civils de Lyon, 69500, Bron, France
3 Équipe GENDEV, Centre de Recherche en Neurosciences de Lyon,INSERM U1028, CNRS UMR5292, UCBL1, 69500, Bron, France
4 Centre de Biotechnologie Cellulaire, Hospices Civils de Lyon, 69500, Bron, France
5 Université de Lyon, Université Lyon 1, CNRS, Laboratoire de Biométrie et Biologie Évolutive UMR 5558, 69622 Villeurbanne, France
6 Service de Biostatistique - Bioinformatique, Hospices Civils de Lyon, 69003, Lyon, France

Auteur Correspondant : pierre-antoine.rollat-farnier@chu-lyon.fr

Le séquençage haut-débit (NGS), utilisé en routine diagnostique pour la caractérisation des variants pathogènes, permet théoriquement de détecter l'ensemble des variants structuraux (SV) d'un patient. En pratique, les stratégies usuelles basées sur la capture de régions cibles (panels de gènes, exomes) s'avèrent inefficaces. Tout d'abord, elles sont inadaptées en ce qui concerne la détection des variants structuraux équilibrés (BCA) tels que les inversions ou les translocations, qui ne laissent d'autres signaux que des points de cassure dont la probabilité d'être capturés est faible. Ainsi, peu de pipelines de routine diagnostique incluent la détection de ces variants. Ensuite, ces techniques induisent des biais de séquençage qui perturbent énormément la détection des variants de nombre de copies (CNV) – c.-à-d. les duplications et délétions.

Le séquençage complet de génomes semble davantage adapté à la détection des SV. Par définition, il permet d'obtenir le génome complet du patient, points de cassure compris. De plus, en s'affranchissant des étapes de capture, il limite les biais susmentionnés. La démocratisation « attendue » de cette stratégie dans les années à venir devrait favoriser la prise en compte des SV en routine diagnostique. Or, si une myriade d'outils de détections des SV est d'ores et déjà disponible, très peu proposent à notre connaissance une annotation pertinente des SV dans un cadre diagnostique.

Pour répondre à cette problématique, nous présentons *Svagga* (pour Structural Variant AGGregation and Annotation), un logiciel écrit en *Perl* et dédié au formatage et à l'annotation des SV. *Svagga* prend en entrées des variants aux formats TAB ou VCF, pour plusieurs échantillons et plusieurs logiciels, ainsi que diverses bases de données aux formats BED et GFF. En premier lieu, *Svagga* va agréger les variants identifiés par plusieurs logiciels, dans le but de résumer l'information pour chaque échantillon. Une comparaison entre les échantillons calcule par la suite les occurrences de ces variants, facilitant par exemple l'élimination des variants fréquents. Enfin, *Svagga* utilise l'information des bases de données fournies afin d'annoter ces variants avec des données biologiques pertinentes (pistes d'UCSC, gènes RefSeq, *etc.*).

Plusieurs paramètres – actionnables par l'utilisateur – vont régir le comportement de *Svagga*. Pour considérer deux SV comme étant identiques, *Svagga* demande par exemple une distance maximale entre leurs points de cassure respectifs, et s'il faut ou non prendre en compte le type d'événement inféré par le logiciel (« délétion » versus « inversion » par exemple). En ce qui concerne les CNV, la notion de chevauchement est primordiale. *Svagga* va également pouvoir jouer sur la réciprocité et la transitivité des relations entre SV, et décider de traiter à part ou non les CNV et les BCA. Enfin, *Svagga* étiquette les SV susceptibles d'appartenir à un même événement complexe, ce qui s'avère utile pour l'étude des chromoanagenesis, caractérisés parfois par des dizaines de points de cassure.

Svagga a déjà pu être testé avec efficacité sur une cinquantaine de génomes dans le cadre du projet de recherche ANI, et est inclus en routine diagnostique dans notre laboratoire de Cytogénétique. Il est disponible en ligne (https://gitlab.inria.fr/NGS/svagga).

French Guiana Severe Syndromes, a metagenomics analysis of unknown dark clinical samples

Hourdel V, Vandenbogaert M, Caro V, Balière C, Bremand L, Labeau B, Moua D, Kwasiborski A, Thiberge JM, Mayence C, Rousset D, Hommel D, Manuguerra JC, Kallel H, Matheus S.

Each year, the intensive care unit of Cayenne hospital in French Guiana reports a number of unresolved and fatal human cases associated with acute febrile severe symptoms (about 15 cases/year). Indeed, this French overseas department located in northeastern of South America has suitable conditions for pathogen emergence, with the expansion of the population, the high rate of deforestation and other ecological changes increasing the risk of zoonotic pathogen transmission to humans.

Therefore, a prospective clinical research program was designed to identify potential pathogens that are responsible for these severe clinical cases reported by the hospital unit using a metagenomics approach based on High Throughput Sequencing (HTS).

Genomic and metagenomics sequencing has already been vital in the identification and characterization of known and novel pathogens. Next generation sequencing (NGS) has gained in throughput and cost-efficiency, strongly affecting public health and biomedical research and enabling the conduct of large-scale genomic projects. In this area of research, metagenomics has become a fast developing field for characterizing microbial communities in environmental and/or clinical samples at the genomic level in order to reach functional and taxonomic conclusions. As such, the accessibility of HTS technologies has thoroughly modified this very field of microbiology, in its capacity to detect pathogens at a level surpassing traditional methods (PCR based). However, some challenges still have to be address before HTS becomes a routinely used tool, especially because of data analysis, which is still lacking standard and easy-to-use protocols driven by robust pipelines. In this context, we developed a pipeline for fast and efficient automatic characterization of microbial genomes from HTS data.

In the current study, 13 patients with different clinical symptoms (hepatic, respiratory or encephalitic failure) were included after first-line negative diagnosis investigations. Then, different biological samples (sera, plasma, urine, cerebrospinal fluid or pharyngeal fluid) depending on their clinical symptoms were collected and subjected to whole genome sequencing performed on an Illumina HiSeq 2500 platform. This setup was used to identify for each included patient the pathogen(s) responsible for his disease and each sample was analyzed individually. The libraries were prepared using TruSeq Nano DNA Library Prep® kit in order to produce paired-end reads of approximately 145 pb each.

In infectious metagenomics studies, it is critical to detect rapidly and with high sensitivity potential life-threatening pathogens. It requires a fast and accurate method for recovering the known microbial biodiversity, but has to be completed with an approach to explore the remaining unknown dark content. A metagenomics pipeline has been implemented to explore the content of each sample. The reads were first trimmed and cleaned, human host sequences were removed and a fast *k-mer* based classification tool was used to obtain taxonomic assignments, which were further refined with more conventional homology-based search methods (blast-type). All samples were then compared to each other to detect potential contaminants.
With k-mer based tools, it is possible to obtain a fast answer about which known microbes are present and their number. The choice of the classifier determines the robustness of the taxonomic and quantitative results. In the context of outbreaks, it has to be time and memory-efficient to enable the analysis of several samples a day. However, it remains crucial to further explore the unknown diversity of metagenomics samples using conventional homology search algorithms. The resulting taxonomic profiles enable thorough sample microbial content description to pinpoint infectious agents, giving leads in the establishment of infectious diagnostics.

With the growing importance of infectious diseases in health care and emerging disease outbreaks, high-throughput technologies and bioinformatics have demonstrated potential to improve public health control of infectious diseases by speeding outbreak detection and response, improving preventive interventions, and detecting emerging infectious diseases.

From primary to tertiary structure analyses of experimentally proven O-GlcNACylated sites for an optimised prediction

Théo Mauri[1], Laurence Menu-Bouaouiche[2], Muriel Bardor[2], Tony Lefebvre[1], Marc Lensink[1]and Guillaume Brysbaert[1]
[1] UGSF, University of Lille, F-59000, Lille, France
[2] Glyco-MEV, UniRouen, F-76000, Rouen, France

Corresponding author: theo.mauri@univ-lille.fr

1 Abstract

The O-GlcNacylation is a PTM (Post Translationnal Modification) which consists in the addition of a UDP-GlcNAc on a serine or threonine. It is catalysed by the OGT (O-GlcNActransferase), while the OGA (O-Glycosylaminase) removes the carbohydrate by hydrolysis [1]. The combined actions of the two makes this PTM reversible and serves to regulate proteins activity [2]. Its deregulation is known to be involved in various diseases like the Alzheimer's disease, cancers and diabetes. [3,4]. The knowledge of the sites that are candidates to O-GlcNAcylation is essential to improve the understanding of this reaction and its impact.

Several software showed up these last years to achieve this goal. They predict O-GlcNAcylated sites based on protein sequences [5,6], but they all show a high level of false positives. We postulated these predictions could be improved with the integration of structural elements in the process.

Thus, we analysed experimentally proven O-GlcNAcylated sites of mammalian proteins from the primary to the tertiary structures and extracted parameters like residue composition, secondary structure and accessibility that will be used for a random forest classification.

Acknowledgements

This work is supported by the CNRS.

References

[1] R. S. Haltiwanger, G. D. Holt, and G. W. Hart. Enzymatic addition of O-GlcNAc to nuclear and cytoplasmic proteins. Identification of a uridine diphospho-N-acetylglucosamine:peptide beta-N-acetylglucosaminyltransferase. *J. Biol. Chem.*, 265(5):2563–2568, February 1990.

[2] G. W. Hart. Dynamic O-linked glycosylation of nuclear and cytoskeletal proteins. *Annu. Rev. Biochem.*, 66:315–335, 1997.

[3] Fei Liu, Khalid Iqbal, Inge Grundke-Iqbal, Gerald W. Hart, and Cheng-Xin Gong. O-GlcNAcylation regulates phosphorylation of tau: a mechanism involved in Alzheimer's disease. *Proc. Natl. Acad. Sci. U.S.A.*, 101(29):10804–10809, July 2004.

[4] Zhiyuan Ma and Keith Vosseller. Cancer metabolism and elevated O-GlcNAc in oncogenic signaling. *J. Biol. Chem.*, 289(50):34457–34465, December 2014.

[5] Hui-Ju Kao, Chien-Hsun Huang, Neil Arvin Bretaña, Cheng-Tsung Lu, Kai-Yao Huang, Shun-Long Weng, and Tzong-Yi Lee. A two-layered machine learning method to identify protein O-GlcNAcylation sites with O-GlcNAc transferase substrate motifs. *BMC Bioinformatics*, 16 Suppl 18:S10, 2015.

[6] Cangzhi Jia, Yun Zuo, and Quan Zou. O-GlcNAcPRED-II: an integrated classification algorithm for identifying O-GlcNAcylation sites based on fuzzy undersampling and a K-means PCA oversampling technique. *Bioinformatics*, 34(12):2029–2036, 2018.

GARDEN-NET: a tool for chromatin 3D interaction network visualization

Miguel MADRID-MENCIA, Vera PANCALDI

Centre de recherche en Cancérologie de Toulouse (CRCT, UMR1037 Inserm / Université Toulouse III Paul Sabatier, 2 Avenue Hubert Curien, 31037, Toulouse, France) & Barcelona Supercomputing Center, Carrer de Jordi Girona, 29-31, 08034, Barcelona, Spain

Corresponding Author: miguel.madrid-mencia@inserm.fr, vera.pancaldi@inserm.fr

1. Introduction

The last few years have seen an explosion in the number of chromosome conformation capture datasets detailing chromatin contacts in multiple species, cell types and cell cycle stages. *Genome ARchitecture DNA Epigenome and Nucleome - Network Exploration Tool* (https://pancaldi.bsc.es/garden-net) is a webtool to interact with this data in a novel way. We represent experimentally identified chromatin 3D contacts as connections between nodes representing genomic fragments and apply network tools to analyse them.

2. Results

GARDEN-NET consists of a clean and minimalist user interface that allows browsing through networks for different organisms and cell types (currently integrating data from more than 10 human haematopoietic cell types [1] and mouse embryonic stem cells [2], facilitating the exploration of the data sets and integration with other genome-wide data. It exploits interfaces to several R packages (ChAseR, Igraph, GenomicRanges, Tidyverse, ...) to go beyond visualization including network analysis tools, integration of the network with other datasets and calculation of properties of chromatin features in relation to the network. For example, it allows for calculation of chromatin assortativity (ChAs) of a feature, which measures whether chromatin fragments with the feature contact each other preferentially in 3D [2]. ChAs has been applied to identify factors associated to chromatin contacts involving promoters [2], and to study the organization of replication in 3D [3] but can be applied on any chromatin interaction dataset in combination with any genomic feature.

The main interface consists of a network viewer (cytoscape.js) in which different properties of the network nodes are mapped to visual parameters including colour, which indicates the presence of the chosen (epi)genomic feature on the chromatin fragment represented by the node. Users can select which chromosome to explore or whether to visualize the genome-wide network. A search bar with automated suggestions allows the user to search the network by gene name or by genomic range leading to the appearance of a zoomed image on the right-most panel. Using the Promoter Capture HiC networks provided, which include interactions involving gene promoters, this allows to explore regulatory regions of specific genes, including their annotation and possible overlap with other genes, as well as to identify genes that are potentially co-transcribed. A table shows the network properties of the selected chromosome (degree, different node and edge categories). The user can also select from a list of features (for example 80 ChIPseq binding profiles in mESC and 6 epigenomic datasets for the human cells) and visualize them on the network. When a feature is selected, a table appears with network parameters of the nodes annotated with the chosen features (average degree, ChAs etc.). Importantly, users can upload their own feature files in a variety of formats and can thus inspect the properties of their feature of interest projected on the chosen network.

This novel visualization framework goes beyond the traditional view of chromatin contact maps as heat-maps and allows the user to explore global as well as local properties of this network, putting their regions of interest in the greater context of 3D chromatin interactions.

References

1. Biola M. Javierre et al. Lineage-specific genome architecture links enhancers and non-coding disease variants to target gene promoters. *Cell* 167(5):1369-1384, 2016.
2. Vera Pancaldi et al. Integrating epigenomic data and 3D genomic structure with a new measure of chromatin assortativity. *Genome Biology* 17:152, 2016.
3. Karolina Jodkowska, Pancaldi et al. Three-dimensional connectivity and chromatin environment mediate the activation efficiency of mammalian DNA replication origins (BioRxiv 644971).

Genetic determinants of intracranial aneurism in autosomal dominant polycystic kidney disease

Claire LEMAN [1,2], Raphaël GAISNE [1,2], Axelle DURAND [1,2], Matilde KARAKACHOFF [4,6], Romain BOURCIER [4,5], Lucile FIGUERES [2,7], Nicolas VINCE [1,2], Richard REDON [4], Hubert DESAL [4,5], Pierre-Antoine GOURRAUD [1,2], Maryvonne HOURMANT [2], Sophie LIMOU [1,2,3]

1 ATIP-Avenir, Centre de Recherche en Transplantation et Immunologie UMR 1064, INSERM, Université de Nantes, Nantes, France
2 Institut de Transplantation Urologie Néphrologie (ITUN), CHU Nantes, Nantes, France
3 École Centrale de Nantes, Nantes, France
4 INSERM, CNRS, Université de Nantes, l'institut du thorax, Nantes, France
5 Department of Neuroradiology, CHU Nantes, Nantes, France
6 Clinique des données, CIC Inserm 1413, CHU Nantes, Nantes, France
7 STEP, Regenerative Medicine and Skeleton UMRS 1229, Inserm, Nantes, France

Corresponding Author: sophie.limou@univ-nantes.fr

Autosomal Dominant Polycystic Kidney Disease (ADPKD) is the most common hereditary kidney disease (prevalence of 1/400 to 1/1000) and the fourth leading cause of end stage renal disease. The most severe extrarenal manifestation of ADPKD is intracranial aneurisms (IA) with a mortality of 40% in case of rupture. IA prevalence in ADPKD is estimated around 9-12%, which is fivefold higher than in the general population. Familial history of IA is the only known risk factor for IA occurrence in ADPKD, suggesting a major role for yet-to-be-identified genetic factors. Here, we aimed to lead the first genomic study of genetic determinants for IA development in ADPKD. [1]

From 500 ADPKD patients followed at the Nantes Hospital, we singled out patients with IA and their relatives. Illumina whole exome sequencing (mean depth of 100X) was performed for ADPKD patients with IA, and for relatives with ADPKD but with no imaging-proven IA as controls. To process the fastq files, we created an analysis workflow with Snakemake using BWA-mem and GATK Haplotype Caller to generate multisample VCFs. We added annotations about genes, functional prediction, pathways, and conservation with ANNOVAR and SnpEff. We validated our pipeline using *Genome in a Bottle* data. We then restricted our dataset to rare variants (MAF<1%) that were predicted deleterious by ≥2 different prediction scores (among CADD, REVEL, VEST3 and FATHMM) and that segregated between cases and controls in the whole cohort first, and then within families. Finally, to maximize our power of detection, we implemented a burden test strategy.

A total of 48 patients with IAs were identified in 40 different pedigrees. The prevalence of IA was 9,6% in ADPKD patients and 23.9% in ADPKD patients with a familial history of IA. For the discovery cohort, 11 patients PKD+IA+ and 11 relatives PKD+IA- were sequenced. 5367 rare variants with good quality had a potential deleterious impact but did not segregate properly between cases and controls. Testing for polygenicity, we looked for causal variants within families, and found a disruptive inframe insertion in *PCNT*, recently associated to IA, in one family. Another variant, in *PFFIA2,* downregulated in IA patients' arteries was present in two independent families. No significant association between IA and our variants aggregated by genes was identified.

IA development in PKD doesn't seem to be caused by a single genetic variant shared by all IA cases, which indicates the polygenicity of IA in PKD. We identified candidate variants in families, that should be confirmed by Sanger sequencing and replication studies.

Key words: Autosomal Dominant Polycystic Kidney disease, Intracranial aneurysm, genetics, whole exome sequencing, Snakemake workflow.

Acknowledgements

This work was supported by grants from "Fondation pour la Recherche Médicale" (FRM M2R201806005900) and PKD-France.

References

1. Perrone RD, Malek AM, Watnick T. Vascular complications in autosomal dominant polycystic kidney disease. *Nat Rev Nephrol.* 2015 Oct;11(10):589–98.

Genome-scale metabolic networks from two asian brown algae : integrating targeted pathways analyses and metabolomic data

Delphine NEGRE[1], Gabriel MARKOV[2] and Erwan CORRE[1]

[1]CNRS, Sorbonne Université, Plateforme ABiMS - Station Biologique de Roscoff (SBR), Place Georges Teissier, 29680, Roscoff, France

[2]CNRS, Sorbonne Université, Laboratoire de Biologie Intégrative des Modèles Marins (LBI2M) - Station Biologique de Roscoff (SBR), Place Georges Teissier, 29680, Roscoff, France

Corresponding Author: delphine.negre@sb-roscoff.fr

Seaweed cultivation is an ancestral practice that appeared in Asian countries and whose first traces were found around the 5th century. Nowadays, the annual global economic value associated with algal production is estimated at $2.5 billion. Considering that edible brown macro-algae such as *S. japonica* represent more than 88% of this global production, understanding the growth mechanisms associated with these organisms is fundamental. As a result, at the species level, studying metabolism through the modeling of metabolic networks is one of the options envisaged to improve our knowledge of algal physiology. For instance, the sporulation of *S. japonica* would be regulated by a phytohormone derived from carotenoids, abscisic acid. However, although the mechanisms related to its biosynthesis and regulation have long been defined in plants, they remain a mystery in algae.

In this perspective, following the approach adopted during the reconstruction of the metabolic network of *Ectocarpus siliculosus* [1], we have reconstructed a first version of the metabolic networks of two brown algae, *Saccharina japonica* and *Cladosiphon okamuranus*, from genome sequencing and annotation data together with the analysis of orthological links with various model organisms. These operations are performed using the Trinotate automatic annotation pipeline and the AuReMe (Automated Reconstruction of Metabolics models) environment [2]. Using such automated tools ensures the traceability of data and the reproducibility of analyses, in accordance with the quality criteria established by the community [3]. These *in silico* steps generally make it possible to reconstruct networks that are qualified as functional, i.e. capable of producing theoretical biomass.

Nevertheless, these methods, although effective, are limited and require manual expertise and knowledge sharing through *collaborative work* to refine the quality and accuracy of these drafts. The completeness of these networks can thus be achieved by adding metabolomic data and studying specific biosynthesis pathways. To illustrate these last points, we have included respectively within the *S. japonica* and *C. okamuranus*' networks 90 and 34 target metabolites described in the literature and we'll present a focus on the biosynthesis pathway of *oxygenated carotenoids*, metabolites produced under stress conditions.

References

1. S. Prigent et al. The genome-scale metabolic network of *Ectocarpus siliculosus* (EctoGEM): a resource to study brown algal physiology and beyond. *The Plant Journal: For Cell and Molecular Biology*, pages 367-381, 2014.
2. M. Aite et al. Traceability, reproducibility and wiki-exploration for "à-la-carte"reconstructions of genome-scale metabolic models, Plos Computational Biology, 2018.
3. I. Thiele et B. Ø. Palsson. A protocol for generating a high-quality genome-scale metabolic reconstruction. *Nature Protocols*, pages 93-121, 2010.

GSAn : Une alternative aux analyses statistiques des groupes de gènes

Aarón Ayllón-Benítez[1,2], Patricia Thébault[2] and Fleur Mougin[1,2]
[1] Univ. Bordeaux, Inserm UMR 1219, Bordeaux Population Health Research Center, ERIAS team, France
[2] Univ. Bordeaux, CNRS UMR 5800, LaBRI, France

Corresponding author: aaron.ayllon-benitez@u-bordeaux.fr

1 INTRODUCTION

La grande masse de données omiques générées grâce aux technologies à haut débit a motivé l'utilisation de stratégies basées sur des méthodes statistiques d'enrichissement pour comprendre les relations entre génotype et phénotype. Nous proposons une approche alternative pour l'annotation de groupes de gènes, appelée GSAn (https://gsan.labri.fr), qui exploite des mesures de similarité sémantique afin de réduire *a priori* l'annotation. L'outil offre une visualisation originale et interactive pour faciliter l'interprétation des résultats par les experts qui peuvent choisir le niveau d'information biologique qui leur semble pertinent.

2 MÉTHODES

GSAn utilise la structure du graphe de la Gene Ontology (GO) et l'annotation fournie pour chaque gène par GO Annotation (GOA). Comme ces annotations sont obtenues de différentes manières (expérimentales ou automatiques), nous supprimons toute annotation redondante ou jugée incomplète.

Les mesures de similarité sémantique permettent de comparer deux termes GO à partir de leurs propriétés (profondeur, contenu d'information, etc.) ou de leurs annotations. Pour chaque terme GO associé aux gènes d'un groupe, une matrice de similarité est constituée et est ensuite fournie en entrée à un algorithme de clustering hiérarchique. Pour chaque cluster de termes obtenu, deux stratégies pour récupérer les termes représentatifs du cluster sont appliquées. Tout d'abord, le ou les termes du cluster qui annotent la majorité des gènes associés au cluster sont recherchés. Si un tel terme n'existe pas, un algorithme de parcours du graphe de la GO est utilisé pour obtenir les termes représentatifs [1].

Finalement, un algorithme basé sur le problème de couverture par ensembles est proposé pour sélectionner les termes les plus synthétiques sans affecter le nombre de gènes couverts par les termes représentatifs.

3 SERVEUR

GSAn permet aux utilisateurs d'annoter une liste de symboles de gènes ou de protéines UNIPROT et fournit un ensemble de visualisations favorisant la compréhension des résultats d'annotation : (i) trois diagrammes en secteurs présentant l'information sur le groupe de gènes (couverture par GOA et GSAn et similarité au sein du groupe), (ii) un diagramme en barres qui montre l'information des termes synthétiques (iii) un tableau qui représente l'information de tous les termes représentatifs et (iv) une combinaison de visualisations arborescentes présentant conjointement les termes représentatifs et les gènes du groupe étudié [2].

4 APPLICATION

Pour illustrer GSAn, nous avons étudié un jeu de données de 360 groupes de gènes issus d'une approche de transcriptomique étudiant la réponse immunitaire dans le cadre d'une étude de vaccination. Deux analyses on été réalisées avec GSAn : (i) une comparaison des résultats d'annotation de GSAn par rapport à des outils d'enrichissement et (ii) l'étude d'un groupe de gènes annoté par des experts comme *régulation de la présentation d'antigènes* et *réponse immunitaire*.

References

[1] Ayllón-Benítez A., Mougin F., Allali J., Thiébaut R., and Thébault P. A new method for evaluating the impacts of semantic similarity measures on the annotation of gene sets. *PLoS ONE*, 2018;13(11):1–22.

[2] Ayllón-Benítez A., Thébault P., Fernández-Breis J.T., Quesada-Martínez M., Mougin F., and Bourqui R. Deciphering gene sets annotations with ontology based visualization. In *Inter. Conf. Info. Vis. (IV)*, 2017.

Hermès : a management tool for Next-Generation Sequencing analysis on a genomic plateforme

Mélissa N'Debi[1], Guillaume Gricourt[1], Vanessa Demontant[1], Anais Nguyen-Goument[1], Abdelrazak Aissat[1,2,4], Paul-Louis Woerther[3] and Christophe Rodriguez[1,2,3]

[1] NGS platform AP-HP - IMRB Institute, Henri Mondor Hospital - University Paris-Est, 94000, Créteil, France

[2] Inserm U955, Henri Mondor Hospital - University Paris-Est, 94000, Créteil, France

[3] Department of Microbiology AP-HP, Henri Mondor Hospital - University Paris-Est, 94000, Créteil, France

[4] Department of Genetics, Henri Mondor Hospital - University Paris-Est, 94000, Créteil, France

Corresponding author: melissa.ndebi@aphp.fr

1 Introduction

The genomic platform of Henri Mondor hospital (Créteil, France) run over 400 next-generation sequencing (NGS) experiments every year. Those are carried out by different hospital and research teams, with particular experimental protocol, sequencing characteristics and bioinformatic analysis. Every steps that occur after the sequencing, are performed manually, including copying data to the server, running bioinformatic analysis with specific parameters, producing a report, saving the results and inform medical staf or researchers that the data are available. These steps, on the bioinformatician control (which is frequently a rare resource in a laboratory), are repetitive and mostly time consuming. To solve these issues, by automating all the process, and making possible to launch them by technicians or engineer, we have developed a new web interface software named "Hermès".

2 Methods

Hermès was developed through the Django framework (version 2.1) in Python 3 [1] and paired with a SQL database to store information about experiments and their analysis. The Technicians and engineers web-interface is available on hospital intranet. After logging with their personal id and password, they fill a very easy form containing information such as operator, sequencer, samples ID, and choose a bioinformatic pipeline. Then, as soon as the data are available, the system perform immediately a copy of data on storage server, launch bioinformatic pipeline, make a report and transfer data results automatically to the recipient. During these operations, the status of the analysis is indicated.

3 Results-Discussion

At this day, this software is implemented with various pipelines in research and clinical diagnostic, have reduced the time to obtained the results, and reduced the time of bioinformaticians. It has also improve the traceability of the sequencing experiments, was judged conform to ISO-EN-NF 15189 Norma and was accredited for medical used (BM MG6 SH FINF 50 v06) in our hospital. Further development are currently in development to automatically generate transverse report generate from data collected with various report to evaluate automate such as sequencers and make possible follow-up according to Diagnostic Norma.

Keywords : NGS, Metagenomic Shotgun, NF EN ISO 15189, Diagnosis, Software, Django, Python, Database

References

[1] Django [computer software]. https://djangoproject.com. Accessed: 2019-03.

Hight-Throughput Sequencing from preservative ethanol and bulk of specimens to jointly assess species and population genetic diversity of colonial ascidians

Marjorie COUTON[1], Aurélien BAUD[1], Claire DAGUIN-THIEBAUT[1], Thierry COMTET[1] and Frédérique VIARD[1]

[1] Sorbonne université, CNRS, UMR 7144, Station Biologique de Roscoff, Place Georges Teissier, 29680 Roscoff, France

Corresponding Author: mcouton@sb-roscoff.fr

Metabarcoding of environmental DNA (eDNA; i.e., DNA extracted from environmental samples such as seawater) or bulk-DNAs (i.e., DNA obtained from a specimen assemblage, such as plankton samples) has been shown to be a valuable tool for ecological researches, especially for biodiversity assessment. Metabarcoding approach starts by the amplification of a small barcode chosen to target and discriminate efficiently the taxa of interest, and is followed by high-throughput sequencing (HTS). Depending on the marker used, taxonomic assignment can be realized at different levels such as species or family. It is reasonable to think that, by targeting a genomic fragment both highly specific and variable, it would be possible to explore intraspecific diversity within a taxon of interest. Examining intraspecific diversity, here haplotypic diversity, is at the heart of population genetic studies, for instance to examine connectivity or demographic events (e.g., founder events). The ability of using eDNA or bulk-DNA metabarcoding to uncover haplotypes has already been successfully tested in either experimental (e.g.,[1]) or field-based settings (e.g., [2]) but questions are still raised about biases inherent to the method, especially sequencing errors or lack of correlation between read counts and haplotype frequencies [3]. In this study, we addressed these concerns by examining two colonial ascidians (*Botrylloides diegensis* and *Botrylloides violaceus*). Being extremely difficult to identify based on morphology only, a portion of the cytochrome oxidase I (COI) is commonly used to distinguish them. We here investigated bulks of the two species: 20 bottles (2 bottles from 10 ports) with preservative ethanol were filled with 15-18 colonies, randomly sampled in the field. Previous to their mixing, a small fragment was collected to barcode each specimen by Sanger sequencing with specific primers targeting a small portion of COI, and thus determine the exact composition (species and haplotype) within each bottle. After 3, 6 and 12 months of preservation in the laboratory, DNA was extracted from 1 mL ethanol from each bottle. In addition, at month 12, the bulk of colonies was crushed and the DNA extracted. COI amplicons were obtained with the same custom-designed primers used for Sanger sequencing. Comparison between the two methods showed that the species as well as the haplotype richness and distribution are well-captured by HTS, and that sequencing errors can be efficiently reduced by an optimized pipeline. When applied to ethanol-based data to evaluate DNA persistence through time in preservation ethanol, the method also allowed determination of species and population diversity. This study thus shows promising results regarding the use of HTS-based approaches for population genetics, including with non-destructive processing of marine specimens preserved in ethanol.

Acknowledgements

The authors are thankful to Diving and Marine core service and their colleagues at Station Biologique of Roscoff, for the sampling. They thank the ABIMS Platform for providing access to calculation resources. This project was supported by the Fondation TOTAL (project Aquanis2.0). MC acknowledges a PhD grant by Région Bretagne and Sorbonne Université.

References

1. Vasco Elbrecht, Ecaterina E Vamos, Dirk Steinke, and Florian Leese. Estimating intraspecific genetic diversity from community DNA metabarcoding data. *PeerJ*, (6): e4644, 2018.
2. Eva E Sigsgaard, Ida B Nielsen, Steffen S Bach, Eline D Lorenzen, David P Robinson, Steen W Knudsen, Mikkel W Pedersen, Mohammed Al Jaidah, Ludovic Orlando, Eske Willerslev, Peter R Møller, and Philip F Thomsen. Population characteristics of a large whale shark aggregation inferred from seawater environmental DNA. *Nature Ecology & Evolution*, (1): 0004, 2016.
3. Clare IM Adams, Michael Knapp, Neil J Gemmell, Gert-Jan Jeunen, Michael Bunce, Miles D Lamare, and Helen R Taylor. Beyond biodiversity: can environmental DNA (eDNA) cut it as a population genetics tool? *Genes*, (10/3): 192, 2019.

How to involve repetitive regions in scaffolding improvement

Rémy Costa[1,4], Quentin Delorme[1,2], Yasmine Mansour[1,2,3], Anna-Sophie Fiston-Lavier[1,3] and Annie Chateau[1,2]

[1] Université de Montpellier

[2] Laboratoire d'Informatique, Robotique et Micro-électronique de Montpellier

[3] Institut des Sciences de l'Évolution de Montpellier

[4] Master Science et Numérique pour la Santé, parcours Bioinformatique, Connaissances, Données

Corresponding author: remy.costa@etu.umontpellier.fr

Abstract

Context and motivation. Repetitive regions (RR) in DNA sequences are present in almost all organisms and may represent over 80% of the genome size. Fundamental source of genetic plasticity and diversity, yet they are a source of complication when it comes to assemble genomes [1]. Assembly produces contigs of various sizes, sometimes really smaller than the original chromosome size. To reduce the fragmentation of chromosomes, the scaffolding process involves additional information, for instance pairing between reads, to infer how contigs are relatively organized [2]. Repetitive regions are disturbing both assembly and scaffolding processes, which are based on graphs. One way to untangle ambiguous parts of these graphs is to use long reads, produced by third-generation sequencing technologies. However, this is not always possible due to high cost and lower quality. Here we propose to use RR sequences themselves to enhance the scaffolding step.

Methodology. The scaffold graph is defined as follows: vertices represent contig extremities, while edges are of two kinds: (1) contig edges, linking both extremities of a contig, and (2) inter-contig edges relating the pairing-information. A weight function on the inter-contig edges indicates how many pairs are supporting this edge. Due to repeats, some of the inter-contigs edges are erroneous and have to be removed from the graph. In other cases, they are supported by RR. Our method is based on a pipeline progressively refining inter-contig edges through RR analysis, described as follows:

1. find the known RR sequences using a repeat database [3], map them on contigs, tag the contigs with this information, and cluster them according to these tags;
2. inside each cluster, determine inter-contig edges sharing coherent RR sequence parts;
3. modify the weight of the validated inter-contig edges;
4. delete edges incoherent with RR composition or length;
5. after scaffolding, use the RR canonical sequence to fill the gaps between contigs.

An additional knowledge about well-documented RRs (such as Transposable Elements) may help to improve Step 2, and answer the following question: do assembly errors come essentially from recent RRs ? Step 3 can be achieved in different ways, thus we propose to try several weight function perturbations. Step 4 is quite expeditious and may be smoothed by introducing a probabilistic measure to ponder the inter-contig weight instead of deleting it.

Validation. The benchmark is composed of organisms offering different repetition rates and sizes. To validate our approach, we use simulated data from model species, amongst them very high quality genomes such as *Drosophila melanogaster* and *Caenorhabditis elegans*. We will carefully examine the influence of each decision step, in the previous pipeline, on the final quality of the scaffolded genome. Also, an analysis will be driven on deleted edges to determine the relevance of this step and calibrate the probabilistic measure. Genome quality will be measured using the QUAST tool [4].

References

[1] Haixu Tang. Genome assembly, rearrangement, and repeats. *Chemical Reviews*, 107(8):3391–3406, 2007.

[2] Martin Hunt, Chris Newbold, Matthew Berriman, and Thomas D. Otto. A comprehensive evaluation of assembly scaffolding tools. *Genome Biology*, 15(3):R42, Mar 2014.

[3] Weidong Bao, Kenji K. Kojima, and Oleksiy Kohany. Repbase Update, a database of repetitive elements in eukaryotic genomes. *Mob DNA*, 6:11, 2015.

[4] Alla Mikheenko, Andrey Prjibelski, Vladislav Saveliev, Dmitry Antipov, and Alexey Gurevich. Versatile genome assembly evaluation with quast-lg. *Bioinformatics*, 34(13):i142–i150, 2018.

IBENS Genomics core facility

Laurent JOURDREN[1], Charlotte BERTHELIER[1], Corinne BLUGEON[1], Fanny COULPIER[1], Karine Dias[1], Bérengère LAFFAY[1,2], Sophie LEMOINE[1] and Stéphane LE CROM[1,3]

[1] Institut de Biologie de l'ENS (IBENS), Département de biologie, École normale supérieure, CNRS, INSERM, Université PSL, 75005 Paris, France

[2] Master Bioinformatique, Normandie Université, UNIROUEN (UNIROUEN), Université de Rouen Normandie, France

[3] Sorbonne Université, CNRS, Institut de Biologie Paris-Seine (IBPS), Laboratory of Computational and Quantitative Biology (LCQB), 75005 Paris, France

Corresponding Author: jourdren@biologie.ens.fr

The **genomics core facility of the Institut de Biologie de l'École normale supérieure** (IBENS) [1,2] was created in 1999. We have been focused on **eukaryotes** and specifically on **functional genomics** analyses since the beginning. We handle classical model organisms and also more exotic organisms (jellyfish, birds, butterflies…). **The facility has always been a well-balanced structure between wet-lab and bioinformatics**: half of the team is involved on the wet-lab part; the remaining half being involved on the data analysis part. Our goal is to help laboratories during their **high-throughput sequencing projects** from the experimental design to data analysis for publication. In 2008, we joined the **France Génomique consortium**, which has been financed by the governmental funding program "Investissement d'Avenir" since 2010. We have been following the **ISO 9001** quality international standard since March 2013 and the NF X 50-900 certification defined by **IBiSA** since April 2015.

All the staff working on the facility gets a balanced schedule between the core **production service** and **research and development projects** to propose **up to date and reliable experimental solutions** to our collaborators. To cope with the experimental constraints of our collaborators among the research teams (a lot of neuroscience and developmental biology teams), we invest a lot of our time in **testing library protocols** (very low quantities, ribosome depletions…). We are also deeply involved in **software development** to manage our project analyses (40% of projects are analysed on the facility). The tools we develop are distributed on an open source basis on **GitHub** [3] and we now provide most of them as **Docker** images [4] to **ease the distribution** of our work. Our concern is to develop workflows to achieve **reproducible and transparent data analysis** of our high throughput experiments.

Since 2016, our facility has been developing two new technologies. The first one is devoted to **single cell RNA-seq** with the buying of a **Chromium** system from **10X Genomics** based on the Drop-seq protocol. The second one is dedicated to **long read** sequencing in RNA-seq. We work with **Oxford Nanopore Technologies MinION** system in order to sequence full length transcripts for isoform abundance estimation. Both technologies are available to our users since 2018.

All these on-going projects allow us to be at the **state of the art in functional genomics** applications so that we can provide the Paris area scientific community all the tools needed to succeed in their high throughput experiments.

References

[1] http://genomique.biologie.ens.fr
[2] Twitter @Genomique_ENS
[3] https://github.com/GenomicParisCentre/
[4] https://hub.docker.com/r/genomicpariscentre/

Identification des proies de gastéropodes venimeux (Conoidea) par approche de métabarcoding

Claire VINCENT[1], Laetitia AZNAR-CORMANO[1], Alexander FEDOSOV[2], Yuri KANTOR[2], Maria-Vittoria MODICA[3], Camille THOMAS-BULLE[1], Nicolas PUILLANDRE[1]

1 Institut Systématique Évolution Biodiversité (ISYEB), Muséum national d'Histoire naturelle, CNRS, Sorbonne Université, EPHE, Université des Antilles, 57 rue Cuvier, CP 26, 75005 Paris, France

2 Institute of Ecology and Evolution, Russian Academy of Sciences, Leninski prospect 33, Moscow 119071, Russia

3 Department EMI, Stazione Zoologica Anton Dohrn, Villa Comunale, I-80121 Naples,Italy

Correspondance : claire.vincent@edu.mnhn.fr

Le succès évolutif des Conoidea, un groupe de gastéropodes venimeux présent dans toutes les mers du monde, et incluant plus de 5000 espèces décrites, serait lié à leur capacité à s'attaquer à de nouvelles proies, via l'acquisition de nouvelles toxines. Pour tester cette hypothèse, il est donc nécessaire d'identifier les proies dont se nourrissent les Conoidea. Celles-ci sont déjà bien caractérisées dans certaines familles telles que les Conidae, mais restent largement inconnues chez les autres familles. Les observations directes étant extrêmement rares, une approche par métabarcoding des contenus stomacaux a été mise au point. L'emploi des NGS dans ce contexte permet de séquencer des échantillons composés potentiellement de plusieurs proies de différentes espèces. Un fragment du marqueur 16S, utilisé comme outil diagnostic chez les annélides, dont se nourrissent préférentiellement la plupart des Conoidea, a été amplifié à partir d'estomacs de plusieurs espèces de trois familles de Conoidea (Turridae, Drilliidae et Pseudomelatomidae). Les données de séquençage obtenues (IonTorrent et Illumina MiSeq) seront ensuite sur une plate-forme en ligne: mBRAVE. Les différents OTUs (Operational Taxonomy Units) identifiés dans les estomacs seront comparés via une approche de reconstruction phylogénétique, de manière à déterminer s'il existe une différence de régime alimentaire entre les espèces des différentes familles.

Identification of a common transcriptional signature for regulatory B cells in Humans and Mice

Florian Dubois[1,2], Sophie Limou[1,2,3], Mélanie Chesneau[1,2], Sophie Brouard[1,2] and Richard Danger[1,2]

1 Centre de Recherche en Transplantation et Immunologie UMR 1064, 30 Bd Jean Monnet, 44093, Nantes, France
2 Institut de Transplantation Urologie Néphrologie (ITUN), CHU Nantes, Nantes, France
3 Ecole Centrale de Nantes, Nantes, France

Corresponding Author: sophie.brouard@univ-nantes.fr

Regulatory B cells (Bregs) have first been described in mice for their ability to regulate inflammation in different model of colitis, EAE and arthritis [1,2,3]. In human, Bregs also demonstrated regulatory function through a variety of mechanisms in different settings [4]. Roles in tolerance mechanisms in kidney and skin transplantation also demonstrate their important role in immunity and their high potential in cell therapy [5,6]. Up to date, no consensual and common Breg phenotype has been described, and whether there is a Breg lineage commitment or if they acquire their function under certain environmental conditions remains unknown.

To address these points, we performed a sample size weighted meta-analysis of publicly available transcriptomic data from 4 different Bregs studies in humans and 5 Bregs studies in mice. Briefly, raw data were processed according a homogenous process in each dataset with differential expression analysis processed between Bregs and non-Bregs. Then a sample size weighted meta-analysis (Stouffer's Z-score method) was conducted using the METAL software [7]. 165 and 126 differentially expressed genes were identified in human and mice respectively with a Bonferroni corrected p-value < 5%. The comparison between humans and mice datasets identified a unique common signature of 4 genes. While we observed high levels of expected genes, such as IL10 in Bregs, we also identified additional genes related to regulatory function in humans and mice, including GZMB and CD9. In order to identify molecules able to discriminate Bregs from non-Bregs, we highlighted 17 genes coding for proteins expressed to the outer side of the cell membrane.

Identification of a unique and common transcriptional Breg signature as well as extracellular markers will allow identifying, characterizing and sorting Breg cells and will offer new options for future cell therapy.

Citations

1. Fillatreau, S., Sweenie, C. H., McGeachy, M. J., Gray, D. & Anderton, S. M. B cells regulate autoimmunity by provision of IL-10. *Nature Immunology* 3, 944–950 (2002).
2. Mizoguchi, A., Mizoguchi, E., Takedatsu, H., Blumberg, R. S. & Bhan, A. K. Chronic Intestinal Inflammatory Condition Generates IL-10-Producing Regulatory B Cell Subset Characterized by CD1d Upregulation. *Immunity* 16, 219–230 (2002).
3. Mauri, C., Gray, D., Mushtaq, N. & Londei, M. Prevention of Arthritis by Interleukin 10–producing B Cells. *The Journal of Experimental Medicine* 197, 489–501 (2003).
4. Oleinika, K., Mauri, C. & Salama, A. D. Effector and regulatory B cells in immune-mediated kidney disease. *Nature Reviews Nephrology* 15, 11–26 (2019).
5. Chesneau, M. *et al.* Unique B Cell Differentiation Profile in Tolerant Kidney Transplant Patients: B Cell Differentiation in Tolerance. *American Journal of Transplantation* 14, 144–155 (2014).
6. Minagawa, R. *et al.* The critical role of Fas-Fas ligand interaction in donor-specific transfusion-induced tolerance to H-Y antigen. *Transplantation* 78, 799–806 (2004).
7. Willer, C. J., Li, Y. & Abecasis, G. R. METAL: fast and efficient meta-analysis of genomewide association scans. *Bioinformatics* 26, 2190–2191 (2010).

Identification of causal signature from omics data integration and network reasoning-based analysis

Méline Wery[1,2], Emmanuelle Becker[1], Franck Augé[2], Charles Bettembourg[2*], Olivier Dameron[1*] and Anne Siegel[1*]
[1] Univ Rennes, Inria, CNRS, IRISA, Rennes, F-35000, France
[2] SANOFI R&D, Translational Sciences, Chilly Mazarin 91385, France
* contributed equivalently to the PhD supervising

Corresponding author: meline.wery@irisa.fr

The identification of a biological signature to diagnose or predict the evolution of a pathology remains a challenge. In most studies, those signatures are patterns of gene expression, used as statistic-based classifiers in order to separate two stratified large populations [1,2]. Furthermore, those sets of selected biomarkers might be used as new targets for drug development.

However, this kind of statistical approach has several limits. First, with a complex and diffuse disease, the clinical criteria are not sufficient to stratificate the population which prevents the gene signature from accurately clustering patients. Second, the identified pattern is a mix between the genes involved in the cause of the pathology, the ones whose expression are induced by the perturbed phenotype and some noise which might be associated with a population heterogeneity. Third, the known interactions (regulation, signal transduction) between the biological entities are not taken into account during the biomarkers identification.

We propose an approach to integrate multi-omics data (genomic and transcriptomic data) and prior knowledge of interaction network in order to compute a causal signature which will be formally defined in a pathological context.

In order to overcome the stratification issue, we used a one-by-one patient approach instead of large population comparison. Each patient is characterized by a set of mutations, and a set of comparisons between the genes expressions and their corresponding interval in the reference population.

We propose to define causal signatures using the principle of minimal intervention sets (MIS) which satisfy a given state of a system, here a biological network [3,4]. In our case, the state of a system is defined by a set of goals (genes expression) and constraints (mutations) for each patient. A causal signature is the minimal set of biological events (MIS) that can explain the gene expression according to mutations. Those signatures can then be used for deriving a novel stratification of the patients and for proposing candidate therapeutic targets. This method will be validated using publically available datasets.

Acknowledgements

This work was supported by SANOFI R&D, Chilly-Mazarin.

References

[1] Bin Xiao, Jianfeng Hang, Ting Lei, Yongyin He, Zhenzhan Kuang, Li Wang, Lidan Chen, Jia He, Weiyun Zhang, Yang Liao, Zhaohui Sun, and Linhai Li. Identification of key genes relevant to the prognosis of ER-positive and ER-negative breast cancer based on a prognostic prediction system. *Molecular Biology Reports*, pages 1–9.

[2] Taixian Li, Yanqiong Zhang, Rongtian Wang, Zhipeng Xue, Shangzhu Li, Yuju Cao, Daobing Liu, Yanfang Niu, Xia Mao, Xiaoyue Wang, Weijie Li, Qiuyan Guo, Minqun Guo, Na Lin, and Weiheng Chen. Discovery and validation an eight-biomarker serum gene signature for the diagnosis of steroid-induced osteonecrosis of the femoral head. *Bone*, 122:199–208, may 2019.

[3] Regina Samaga, Axel Von Kamp, and Steffen Klamt. Computing Combinatorial Intervention Strategies and Failure Modes in Signaling Networks. *Journal of Computational Biology*, 17(1):39–53, 2010.

[4] Santiago Videla, Julio Saez-Rodriguez, Carito Guziolowski, and Anne Siegel. caspo: a toolbox for automated reasoning on the response of logical signaling networks families. *Bioinformatics (Oxford, England)*, 33(6):947–950, mar 2017.

Identification of genomic regions for high-resolution taxonomic profiling using long-read sequencing technology

Jean MAINGUY[1], Olivier BOUCHEZ[2], Adrien CASTINEL[2], Sylvie COMBES[3], Christine GASPIN[1], Denis MILAN[2], Cécile DONNADIEU[2], Carole IAMPIETRO[2], Claire HOEDE[1] and Géraldine PASCAL[3]

1 MIAT, PF Bioinfo GenoToul, Université de Toulouse, INRA, Chemin de Borde Rouge, 31320 Castanet-Tolosan, France
2 INRA, US 1426, GeT-PlaGe, Genotoul, 31320 Castanet-Tolosan, France
3 GenPhySE, Université de Toulouse, INRA, INPT, ENVT, Chemin de Borde Rouge, 31320 Castanet Tolosan, France

Corresponding Author: geraldine.pascal@inra.fr

Taxonomic profiling of microbiome is a challenging task. The 16S rRNA gene is the most used marker to address this question as it is universally distributed among prokaryotes and has conserved and hypervariable regions. On top of that, 16S databases contain genes from many more species than genome databases making them more comprehensive. Studies usually target a small part of the 16S gene depending on the lineage of interest and sequence it with the MiSeq Illumina sequencer. However, the 16S often fails to assign taxonomy at the genus or species level because amplicons are not specific enough and because of the sampling bias of the species contained in the databases. Moreover, with 16S rRNA genes, it is difficult to have a resolving quantitative estimation of different species because this marker is often present in multicopies and we know rarely the number of these copies by species that can be non-cultivable. Alternative markers such as rpoB, gyrB and recA show better results within specific lineage [1]. In parallel, long amplicon PCR-based approaches targeting the full-length 16S rRNA and the whole ribosomal RNA operon show encouraging results [2-4].

We focus here on identification of alternative genomic regions that could be used by long-read sequencing approaches to get a more specific taxonomic resolution. Long read technology enables the sequencing of an entire gene or groups of genes. A longer sequence may bring more information to discriminate closely related organisms. Also, this major advantage may help to offset the high error rate associated with long read technologies. The genomic regions that we identify as possible markers consist of single copy and universally distributed genes. We investigate the possibility to target genomic regions bounded by two of these genes, as universal as possible. At first, to select potential marker genes, we use orthologous groups from eggNOG 5.0 [5] made up of single copy and universally distributed genes from bacteria. We get genes from 5500 refseq representative genomes to find their genomic positions in order to establish relative distances within each genome. Finally, we select potential genomic regions based on the variation of this distance and their discriminating power.

We are implementing the workflow to identify potential genomic regions of interests with Nextflow [6]. It allows having a reproducible and generic framework. Our workflow is then not limited to bacteria but can be used to establish the most adequate genomic regions within any specific lineage or group of lineages.

Acknowledgements

This work is part of the SeqOccIN project supported by Region Occitanie and FEDER

References

[1] Poirier, S. et al. Deciphering intra-species bacterial diversity of meat and seafood spoilage microbiota using gyrB amplicon sequencing: A comparative analysis with 16S rDNA V3-V4 amplicon sequencing. PloS one, 13(9), e0204629.
[2] Shin, J et al. Analysis of the mouse gut microbiome using full-length 16S rRNA amplicon sequencing. Scientific reports, 6, 29681.
[3] Benítez-Páez, A., & Sanz, Y. (2017). Multi-locus and long amplicon sequencing approach to study microbial diversity at species level using the MinION™ portable nanopore sequencer. Gigascience, 6(7), gix043.
[4] Cusco, A. et al. Microbiota profiling with long amplicons using Nanopore sequencing: full-length 16S rRNA gene and whole rrn operon. F1000Research, 7.
[5] Huerta-Cepas et al. eggNOG 5.0: a hierarchical, functionally and phylogenetically annotated orthology resource based on 5090 organisms and 2502 viruses. Nucleic acids research, 47(D1), D309-D314.
[6] Paolo Di Tommaso et al. Nextflow enables reproducible computational workflows. *Nat Biotechnol.*, 35:316-319, 2017.

Identifying predictive biomarkers for breast cancer treatment using an integrative transcriptomic analysis

Agnes BASSEVILLE[1], Fabien PANLOUP[2], Bertrand MICHEL[3] and Philippe JUIN[1]

[1] CRCINA, 8 quai Moncousu, BP 70721, 44007 NANTES Cedex 01, France

[2] LAREMA, Faculté des Sciences, 2 Boulevard Lavoisier, 49045 Angers cedex 01, France

[3] Laboratoire de Mathématiques Jean Leray, 2 Rue de la Houssinière, BP 92208, 44322 Nantes Cedex 03, France

Corresponding Author: agnes.basseville@inserm.fr

Breast cancer is the cancer with the highest incidence in women worldwide, and is the main cause of cancer-related death for women [1]. This high death rate is mainly due to the lack of treatment for metastatic breast cancer, the existence of breast cancer subtypes refractory to treatment, and the high level of tumor heterogeneity (defining the very diverse cell populations composing each tumor) leading to therapy failure. Breast cancer subtypes and tumor heterogeneity suggest the need for the development of tailored, personalized treatments, but so far, the discovery of efficient predictive markers has been compromised by the lack of adapted biological models and methodological tools [2]. With recent developments of high-throughput methods in biology, large 'omics' datasets from patients now offer a better understanding of tumor complexity. These biomedical 'omics' are typically small-sample size with high dimensions of features, which implies the use of adapted mathematical tools to identify important features [3]. Nevertheless, no consensus methodology has arisen yet in the search for biomarkers from omics data [4].

In this work, we developed a pipeline to define predictive biomarkers for breast cancer therapy using transcriptomic data from the open access database GEO. We selected microarrays obtained from patients prior to treatment for whom chemotherapy response was clinically determined afterward. First, we combined 13 microarray datasets to grant a sufficient statistical power (2284 patients) to reveal a comprehensive overview of tumor complexity. Then, we compared the performances of various machine learning algorithms (PLS, LASSO, elastic net, random forest, XGBoost) using prediction error from cross-validation, and confusion matrix. We also tested the performance of our model in mega-analysis versus meta-analysis, with/without feature pre-selection, and with/without resampling strategies. Important variables were then investigated by pathway analysis in order to extract the biological meaning of these biomarkers.

This study identified candidate gene signatures to predict treatment response for breast cancer patients, with a minimum 15-20 % prediction error. Our study confirms the importance of metric choice for algorithm predictions in class-unbalanced datasets and advocates for the use of resampling strategies. Mega-analysis of sufficiently large biological or platform subgroups resulted in the most accurate prediction rate.

References

1. World Health Organization website (http://www.who.int/cancer/detection/breastcancer/en/)
2. Harris EER. Precision Medicine for Breast Cancer: The Paths to Truly Individualized Diagnosis and Treatment. *Int J Breast Cancer*,2018:4809183, 2018.
3. Iain M. Johnstone and D. Michael Titterington. Statistical challenges of high-dimensional data. *Philos Trans A Math Phys Eng Sci.,* 367(1906): 4237–4253, 2009.
4. Koscielny S. Why most gene expression signatures of tumors have not been useful in the clinic. *Sci Transl Med.,* 2(14):14ps2. 2010.

Ignoring the optimal set of tissue-specific metabolic networks can bias the intepretation of data

Pablo Rodriguez*[1], Nathalie Poupin[1], and Fabien Jourdan[1]

[1]INRA Toxalim – Institut national de la recherche agronomique (INRA) : UMR1331 – France

Résumé

Tissue specific constraint-based modeling approachess have proven useful as automatic ways of extracting and analyzing metabolic networks that capture the different metabolic states of cells. These methods integrate different sources of information such as stoichiometry, transcriptomics, metabolomics or fluxes that constrain the space of possible networks that to better describe the metabolic state of cells for a given condition. This process is usually done by searching for a metabolic network that minimizes an objective function measuring the discrepancy between the observed data and the model. However, current methods usually extract one single optimal network from which all the subsequent analysis and interpretation is derived. But depending on the method and data used, this solution may be not unique, meaning that the observed data can be explained by a set of equally good metabolic networks, representing slightly different hypothesis of the metabolic state. Ignoring this variability may lead to incorrect or incomplete explanations and bias subsequent analysis. In order to analyze the impact that this optimal set of networks can have in the interpretation of results, we developed an extension of iMAT, a method for extracting tissue-specific networks from transcriptomics data, to enumerate different alternative optimal networks generated from the same data. Our study highlights the importance of analyzing the space of alternative optimal solutions as a way to reduce potential bias in the interpretation of data using constraint-based modeling approaches.

*Intervenant

Impact de la manipulation thermique embryonnaire sur le méthylome de caille japonaise

Coralie Gimonnet*[1], Anaïs Vitorino Carvalho[1], Nathalie Couroussé[1], Sabine Crochet[1], Thierry Bordeau[1], Marjorie Mersch[2], Benoît Piégu[3], Christelle Hennequet-Antier[1], Aurélien Brionne[1], Frédérique Pitel[2], Anne Collin[1], and Vincent Coustham[1]

[1]BOA, INRA, Université de Tours, 37380 Nouzilly – Institut National de la Recherche Agronomique - INRA – France

[2]GenPhySE, Université de Toulouse, INRA, INPT, INP-ENVT, 31326 Castanet-Tolosan, France – Institut national de la recherche agronomique (INRA) – France

[3]PRC, CNRS, IFCE, INRA, Université de Tours, 37380 Nouzilly – Institut national de la recherche agronomique (INRA) – France

Résumé

Des changements environnementaux pendant l'embryogenèse peuvent avoir un impact sur le développement et le phénotype de l'individu. Dans ce contexte, la thermo-manipulation embryonnaire (TM) chez la caille, consistant en une augmentation cyclique de la température des œufs (1,7∘C) entre les jours 0 et 13, a été étudiée à 0 et 35 jours post-éclosion. Nous avons émis l'hypothèse que la TM pourrait impacter l'expression des gènes *via* des reprogrammations épigénétiques mises en place pendant l'incubation et persistant au cours du développement.
Pour cela, nous nous sommes intéressés à la méthylation de l'ADN dans l'hypothalamus (tissu impliqué dans la thermorégulation au niveau central) de cailles mâles et femelles issues d'une lignée consanguine. Afin de contrôler la variabilité génétique, nous avons utilisé pour chaque répétition biologique une caille contrôle et une caille TM issues du même parent.

La distribution de la méthylation de l'ADN sur l'ensemble du génome a été obtenue par Whole Genome Bisulfite Sequencing avec un séquenceur Illumina NovaSeq 6000 pour une profondeur de 30X. Les lectures pairées de 150 bp obtenues ont alors été traitées par un pipeline que nous avons développé à l'aide du gestionnaire Nextflow permettant la reproductibilité des données. Les lectures ont été cartographiées sur le génome de référence de la caille (Coturnix japonica 2.0, NBCI) composé de 33 chromosomes et 1979 scaffolds et d'une taille d'environ 1Gb. Le pipeline développé permet le traitement automatique des séquences brutes jusqu'à l'obtention des positions des CpG méthylées exploitables pour l'analyse différentielle tout en passant par différentes étapes de nettoyage des données. La recherche de la méthylation différentielle des CpG de l'ADN a ensuite été réalisée avec le package R DSS [1] en prenant en compte l'appariement des données dû à l'effet fratrie.

En plus de cette analyse de la méthylation différentielle entre cailles TM et contrôle, nous avons profiter de la disponibilité des échantillons de chaque sexe pour effectuer une analyse de la méthylation différentielle des mâles *versus* femelles. Cette analyse a été réalisée dans

*Intervenant

le but de valider le pipeline utilisé.
L'ensemble des analyses réalisées dans ce contexte a permis de mettre en place un pipeline utilisable pour d'autres données WGBS et en cours d'adaptation pour permettre également une analyse de données RRBS.

In-silico benchmark of methods for detecting differentially abundant features between metagenomics samples

Léonard Dubois*[1], Magali Berland[1], and Mahendra Mariadassou[2]

[1]MetaGenoPolis – Institut national de la recherche agronomique [Jouy en Josas] – France
[2]MaIAGE – Institut national de la recherche agronomique (INRA), Institut National de la Recherche Agronomique - INRA – France

Résumé

Metagenomics studies microbial communities by sequencing their genetic material. This is done by targeting either a marker-gene (barcoding) or all the genes present in the samples (shotgun sequencing). It has been extensively used to characterize taxonomic and functional profiles of many ecosystems including the human gut microbiota. In the later, the shifts in the abundancy of specific species are used as biomarkers for many biological or clinical conditions such as cancer, diabetes or inflammatory bowel disease. However a lot of technical difficulties arise when working with such data: high-dimension, noise, high sparsity, low number of replicates... Thus, detecting these shifts with satisfying precision and recall is challenging. Many statistical methods were introduced in the past decade, each trying to overcome specific constraint of the data. In this study we present a benchmark of the most commonly used methods for detecting differentially abundant features between samples. By using simulated data, statistical performances such as true/false positive or negative rates are assessed. In addition, results on real microbiome datasets are qualitatively discussed.

*Intervenant

Industrial NGS analysis processes from sequencing to variant interpretation on MOABI platform

Jocelyn BRAYET[1], Camille BARETTE[2], Mathieu BARTHELEMY[1], Romain DAVEAU[1], Vivien DESHAIES[1], Laurent FROBERT[1], and Alban LERMINE[1,2]

[1] MOABI (Bioinformatic platform of AP-HP), 33 boulevard Picpus, 75012, Paris, France
[2] SeqOIA-IT, 33 boulevard Picpus, 75012, Paris, France

Corresponding Author: jocelyn.brayet@aphp.fr

The Assistance Publique – Hôpitaux de Paris (AP-HP) is a teaching hospital groupment with a European dimension globally recognized. The AP-HP is organized into twelve hospital groups, for a total of 39 hospitals localized in Paris and its region. Currently, those hospitals attend each year 8 millions patients.

Two years ago, MOABI, a new bioinformatics platform was created for multiple missions: the progressive storage centralization for genomic data routinely produced by hospitals, their analyses in controlled and standardized workflows and the provisioning of tools for results exploitation.

In this abstract, we present two softwares designed to analyze NGS diagnosis data: G-route and Leaves. The first tool is a java n-tier rich client web application that provides to users raw sequencing data loading, traceability metadata definition, analysis pipelines running and data files browsing. Data files management relies on the adaptative middleware iRODS [1]. At the present time, G-route contains 4250 analyzed patients, 131 users, 11 skeletons and 22 versions of pipelines for 55 different gene panels. This makes it possible to propose an offer of 100 different pipeline combinations. The second program, Leaves, is an open source tool that aim to help biologists for genetic alterations interpretation and biological report generation by associating detected alterations with different annotations and scores and performing reproducible filters combination and ranking. Leaves is a web interface mainly developed with python 3 and javascript. Currently, Leaves's database contains 128 users, 50 projects, 2.041.926 variants and 70 variants classifications. Leaves also permit the sharing of AP-HP expertise, in a standard way, between biologists, promoting human interaction over artificial intelligence. Users can run analyses from G-route through 100 different pipelines that end up inserting variants calling results into Leaves. Pipelines are written in Snakemake [2], that use Docker [3] containers as version fixed tools. Docker allows to eliminate tool dependencies problems and sets a version tool in an image. Presently, we have more than 100 tools integrated in that way. Snakemake is a workflow management system with implicit rule implementation (input and output logic). The advantage over Nextflow [4] is the capability to share rules between pipelines which allowed us to create a rule library (137 rules) that can be shared between pipelines. As other pipeline frameworks, error recovery, automatic parallelization and workflow integrity features are included in Snakemake. To ease medical diagnosis routine, AP-HP's scientists are able to execute tagged workflows with their data and to consult results through G-route and Leaves interfaces.

Key words: Industrial NGS analysis, variants interpretation, iRODS, Docker and Snakemake

References

[1] Web site: https://irods.org/
[2] Johannes Köster, Sven Rahmann; Snakemake—a scalable bioinformatics workflow engine. *Bioinformatics* 2012; 28 (19): 2520-2522. doi: 10.1093/bioinformatics/bts480
[3] Web site: https://www.docker.com/
[4] Jeremy Leipzig; A review of bioinformatic pipeline frameworks. *Brief Bioinform* 2017; 18 (3): 530-536. doi: 10.1093/bib/bbw020

INEX-MED: INtegration and EXploration of heterogeneous bio-MEDical data

Kirsley Chennen[1,2,3,⋆], Maxime Folschette[1,4,⋆], Alban Gaignard[5], Richard Redon[5], Hala Skaf-Molli[4], Olivier Poch[2], Jocelyn Laporte[3], Julie Thompson[2] and the INEX-MED consortium

[1] Institut Français de Bioinformatique, CNRS UMS 3601, France
[2] CSTB - iCUBE, CNRS UMR 7357, Faculté de Médecine, Strasbourg, France
[3] Institut de Génétique et de Biologie Moléculaire et Cellulaire (IGBMC), INSERM U1258, CNRS UMR7104, Illkirch, France
[4] LS2N, Laboratoire des Sciences du Numérique de Nantes, CNRS UMR 6004, Université de Nantes, France
[5] Institut du Thorax, Inserm UMR 1087, CNRS UMR 6291, Université de Nantes, France

Corresponding author: kchennen@unistra.fr, maxime.folschette@ls2n.fr, alban.gaignard@univ-nantes.fr

Abstract

The new era of modern biology and medicine calls for the development of novel integrated approaches leveraging massive multi-disciplinary, multi-scale, and multi-modal biological data. These clinical, imaging, or "omic" datasets are currently stored in data silos [1] which makes their cross-exploitation challenging.

In this poster, we present INEX-MED, a unified Knowledge-Graph based framework aimed at linking diverse data modalities and clinical observations to accelerate both the statistical/semantic data exploitation and its reuse. Following the "FAIR" data principles (Findability, Accessibility, Interoperability, Reusability) [2], we propose a prototype system which (i) integrates clinical, imaging and genomics data from cohorts into a dedicated knowledge-graph, (ii) allows secure access and query processing on heterogeneous biomedical data, and (iii) performs statistical analysis and machine learning to improve the diagnosis/prognosis of studied diseases.

The INEX-MED prototype is currently being developed for two use-case cohorts:

(i) The ICAN cohort [3] covering 3000 individuals affected by intracranial aneurysm for which clinical records, MRI imaging and exome sequencing data are acquired. The aim is to identify biomarkers and risk factors characterising the development of this disorder.

(ii) The MYO-lico cohort with 1200 congenital myopathy patients having clinical records, histopathological imaging data, and exome sequencing data. The aim is to identify novel genes causing congenital myopathies and classify them [4].

INEX-MED already benefits from the Cloud infrastructure provided by the French Bioinformatics Institute (IFB). As a result of this project, the developed prototype will be made available to the biomedical community as an IFB resource, providing methods and technological guidelines to address other biomedical use cases.

References

[1] Akram Alyass, Michelle Turcotte, and David Meyre. From big data analysis to personalized medicine for all: challenges and opportunities. *BMC Medical Genomics*, 8(1):33, Jun 2015.

[2] Mark D Wilkinson, Michel Dumontier, IJsbrand Jan Aalbersberg, Gabrielle Appleton, Myles Axton, Arie Baak, Niklas Blomberg, Jan-Willem Boiten, Luiz Bonino da Silva Santos, Philip E Bourne, et al. The FAIR Guiding Principles for scientific data management and stewardship. *Scientific data*, 3, 2016.

[3] Romain Bourcier, Stéphanie Chatel, Emmanuelle Bourcereau, Solène Jouan, Hervé Le Marec, Benjamin Daumas-Duport, Mathieu Sevin-Allouet, Benoit Guillon, Vincent Roualdes, Tanguy Riem, et al. Understanding the pathophysiology of intracranial aneurysm: The ICAN Project. *Neurosurgery*, 80(4):621–626, 2017.

[4] J Böhm, R Schneider, E Malfatti, V Schartner, X Lornage, I Nelson, G Bonne, B Eymard, J Nectoux, F Leturcq, et al. Integrated analysis of the large-scale sequencing project "Myocapture" to identify novel genes for myopathies. *Neuromuscular Disorders*, 27:S195, 2017.

⋆. These authors contributed equally to this work

Integration of transcriptomic and proteomic data for biomarker discovery in Lassa fever

Emeline PERTHAME[1] and Natalia PIETROSEMOLI[1]
[1]Hub de Bioinformatique et Biostatistique - C3BI, Institut Pasteur, USR 3756 CNRS, Paris, France

Corresponding author: natalia.pietrosemoli@pasteur.fr

The post-genomic era is quickly converging towards the multi-omics era, where high-throughput technologies are exploited to a point that multiple and heterogeneous measurements are more and more possible on the same biological samples. This offers unprecedented opportunities for disentangling the molecular mechanisms of biological systems, yet requires the development of sound statistical and bioinformatic tools for the integration and interpretation of diverse large-scale omics data.

We present a case study consisting on transcriptomic and proteomic data provided by the Unité de Biologie des Infections Virales Emergentes, Institut Pasteur, Lyon, France. The challenge of 3 vaccine candidates against Lassa fever was performed on 12 males macacas (4 replicates per condition). In the experiment, peripheral blood mononuclear cell (PBMC) and plasma were extracted at 6 time points from the day of immunization to 2 weeks after immunization. Total RNA was extracted from PBMC and gene expression was quantified using RNA-Seq technology. The transcriptomic dataset consisted in read counts for 19469 genes. Protein abundance was quantified from plasma samples using tandem mass spectrometry (MS/MS). The proteomic dataset consisted in MS/MS spectra assignment counts for 350 identified proteins. Because of their close biological relationship, gene expression and protein abundance datasets represent great candidates for testing data integration methods aimed at identifying markers distinguishing the three vaccines. Interestingly, despite this close biological relation, the observed correlation among gene expression and protein abundance is often very low [1] at gene/protein level. Thus, more elaborated statistical methods such as data integration methods are required.

We compared an approach based on correlation analysis of the two data blocks (i.e. proteomic and transcriptomic) with an approach taking into account the biological functions of each block. Regularized Generalized Canonical Correlation Analysis for Multiblock Data (RGCCA)[2,3] and its sparse extension SGCCA [4] were used to study the relationship between the two data blocks and to select unique features within each block (e.g. protein/gene markers). Moreover, the Competitive Gene Set Test Accounting for Inter-gene Correlation method [5] was applied to identify molecular pathways consisting in gene/proteins sets that were differentially expressed in each block. Our work shows that even if transcriptomic and proteomic datasets may not have direct overlap, at the functional context they might refer to the same biological pathways or biological processes.

Acknowledgements

We thank Sylvain Baize, Nicolas Baillet and Mathieu Mateo (Unité de Biologie des Infections Virales Emergentes, Institut Pasteur, Lyon, France) for providing us the data used for this application study.

References

[1] S. Haider and R. Pal. Integrated analysis of transcriptomic and proteomic data. *Current Genomics*, (14):91–110, 2013.

[2] A. Tenenhaus and M. Tenenhaus. Regularized generalized canonical correlation analysis. *Psychometrika*, (76(2)):257–284, 2011.

[3] M. Tenenhaus, A. Tenenhaus, and PJF. Groenen. Regularized generalized canonical correlation analysis: a framework for sequential multiblock component methods. *Psychometrika*, (82(3)):737–777, 2017.

[4] V. Guillemot et al. A. Tenenhaus, C. Philippe. Variable selection for generalized canonical correlation analysis. *Biostatistics*, (15(3)):569–83, 2014.

[5] D. Wu and G.K. Smyth. Camera: a competitive gene set test accounting for inter-gene correlation. *Nucleic Acids Research*, (40(17)):e133, 2012.

Interactions de SNPs d'ordre N par pattern mining

Gwendal Virlet[1], Erwan Corlouer[2], Alexandre Termier[1], Anne Laperche[2], Nathalie Nesi[2] et Dominique Lavenier[1]

[1] Irisa, Avenue du général Leclerc, 35000, Rennes, France
[2] IGEPP, Domaine de la Motte, 35653, Le Rheu, France

Auteur référent: gwendal.virlet@irisa.fr

Les technologies de séquençage associées à des traitements bio-informatiques permettent de détecter des variations génétiques, de type SNP (Single Nucleotide Polymorphism) entre individus d'une même population. Ces variations génétiques peuvent être à l'origine de différents phénotypes chez des individus, par exemple une maladie [1] chez l'humain ou une différence de caractère chez la plante [2]. L'étude d'association (Genome wide association study ou GWAS) est une méthode permettant d'identifier des marqueurs ayant un lien avec cette variation de phénotype. Cependant, les SNPs n'expliquent souvent pas toute l'héritabilité [3,4]. L'épistasie peut expliquer cette héritabilité manquante.

Une nouvelle méthode de détection d'interactions d'ordre N a été développée par les équipes GenScale et Lacodam INRIA/IRISA, Rennes. Elle se base sur des techniques de pattern mining et est implémentée dans le logiciel SSDPS (statisticaly significant discriminative pattern search) [5]. Ce logiciel sélectionne des combinaisons de SNPs à partir de deux populations : l'une possédant le trait et l'autre pas. Puis une analyse statistique MB_MDR (model-based multifactor dimensionality reduction) [6] est appliquée pour vérifier la présence d'interactions. Cette approche permet d'étudier des interactions d'ordre n sans a priori sur l'existence d'une interaction d'ordre n-1 contrairement aux autres méthodes existantes et ceci sur des SNPs en équilibre de liaison et sur une taille d'échantillon moyenne (environ 100 individus). Les simulations montrent une puissance de détection d'interaction d'ordre 2 plus faible que ses concurrents mais la méthode semble éliminer la plupart des faux positifs sur les interactions d'ordre supérieur.

Cette méthode a été testée sur une centaine de variétés de Colza relativement à leur date de floraison. Deux populations ont été constituées, une en floraison précoce et l'autre en floraison tardive. 30000 SNPs ont été analysés avec SSDPS puis filtrés par MB_MDR. Le résultat est une liste de 17 patterns de SNPs localisés dans une quinzaine de gènes dont les termes 'flower' ou 'flowering' de l'ontologie 'plant ontology' sont très souvent retrouvés.

Remerciements

Ce travail est soutenu par le PIA Rapsodyn.

Références

[1] Peter M Visscher, Naomi R Wray, Qian Zhang, Pamela Sklar, Mark I McCarthy, Matthew A Brown, and Jian Yang. 10 years of gwas discovery : biology, function, and translation. *The American Journal of Human Genetics*, 101(1) :5–22, 2017.

[2] Matteo Togninalli, Ümit Seren, Dazhe Meng, Joffrey Fitz, Magnus Nordborg, Detlef Weigel, Karsten Borgwardt, Arthur Korte, and Dominik G Grimm. The aragwas catalog : a curated and standardized arabidopsis thaliana gwas catalog. *Nucleic acids research*, 46(D1) :D1150–D1156, 2017.

[3] Brendan Maher. Personal genomes : The case of the missing heritability. *Nature News*, 456(7218) :18–21, 2008.

[4] Wen-Hua Wei, Gibran Hemani, and Chris S Haley. Detecting epistasis in human complex traits. *Nature Reviews Genetics*, 15(11) :722, 2014.

[5] Hoang Son Pham. *Novel Pattern Mining Techniques for Genome-wide Association Studies. Bioinformatics [q-bio.QM].* PhD thesis, IRISA, equipe GENSCALE, 2017.

[6] M Luz Calle, V Urrea, Gemma Vellalta, N Malats, and KV Steen. Improving strategies for detecting genetic patterns of disease susceptibility in association studies. *Statistics in medicine*, 27(30) :6532–6546, 2008.

Joint analysis of multiple compositional data

Antoine Menard[1], Pascale Vonaesch[2], Emna Achouri[1], Hervé Abdi[3] and Vincent Guillemot[1]
[1] Hub de Bioinformatique et Biostatistique - C3BI, Institut Pasteur, USR 3756 CNRS, Paris, France
[2] Molecular Microbial Pathogenesis Unit - Institut Pasteur, U786 INSERM, Paris, France
[3] School of Behavioral and Brain Sciences, The University of Texas at Dallas, Richardson, TX, USA

Corresponding author: vincent.guillemot@pasteur.fr

Afribiota is an international, multi-center project aiming at unraveling pathophysiological changes associated with growth delay (stunting) in children aged 2-5 years on several sources of data [1]. Globally, one in four children under the age of five is suffering from growth delay. To date, the exact pathophysiological mechanisms underlying stunting remain unclear, but several factors are suspected of contributing to the syndrome, including the intestinal microbiota, specific metabolites as well as nutritional deficiencies. These entities are all tightly linked, calling for multidimensional and multiblock data analysis approaches.

In the context of Afribiota, two compositional datasets were collected on two to five-year-old children from Madagascar and Bangui ($n = 926$): gut microbiota composition and relative abundances of gut metabolites (more specifically bile acids and derivatives). Along with these two blocks of data, socio-economic and clinical parameters were obtained through an Electronic Case Report Form. Compositional data are a particular case of multivariate data where the sum of all the values for an individual is always equal to a fixed scalar value, typically 1 or 100. Classical summary statistics on such data (such as standard deviation or mean) vary according to the set of variables of interest. Using traditional statistics on compositional data is therefore not recommended because their interpretation is subject to the variables on which the analysis is focused [2]. A solution for analyzing compositional data is to work with log-ratios of compositions, such as in the Log Ratio Analysis (LRA) [2].

We present in this poster an extension of the LRA to the joint analysis of one or more blocks of compositional data, with two different multiblock methods: Regularized Generalized Canonical Correlation Analysis [3] and Multiple Factor Analysis [4].

References

[1] Pascale Vonaesch, Rindra Randremanana, Jean-Chrysostome Gody, Jean-Marc Collard, Tamara Giles-Vernick, Maria Doria, Inès Vigan-Womas, Pierre-Alain Rubbo, Aurélie Etienne, Emilson Jean Andriatahirintsoa, Nathalie Kapel, Eric Brown, Kelsey E. Huus, Darragh Duffy, B.Brett Finlay, Milena Hasan, Francis Allen Hunald, Annick Robinson, Alexandre Manirakiza, Laura Wegener-Parfrey, Muriel Vray, and Philippe J. Sansonetti. Identifying the etiology and pathophysiology underlying stunting and environmental enteropathy: study protocol of the AFRIBIOTA project. *BMC Pediatrics*, 18(1):236, dec 2018.

[2] Michael Greenacre. *Compositional Data Analysis in Practice*. Chapman & Hall/CRC, August 2018.

[3] Arthur Tenenhaus and Michel Tenenhaus. Regularized Generalized Canonical Correlation Analysis. *Psychometrika*, 76(2):257–284, mar 2011.

[4] Hervé Abdi, Lynne J. Williams, and Domininique Valentin. Multiple factor analysis: principal component analysis for multitable and multiblock data sets: Multiple factor analysis. *Wiley Interdisciplinary Reviews: Computational Statistics*, 5(2):149–179, March 2013.

Large-scale RNA-seq datasets enable the detection of genes with a differential expression dispersion in cancer

Christophe Le Priol[1], Chloé-Agathe Azencott[2,3,4] and Xavier Gidrol[1]
[1] Univ. Grenoble Alpes, CEA, Inserm, BIG-BGE, 38000, Grenoble, France
[2] Center for Computational Biology, Mines ParisTech, PSL Research University, Paris, France
[3] Institut Curie, F-75248, Paris, France
[4] INSERM U900, F-75248, Paris, France

Corresponding author: christophe.lepriol@cea.fr

The majority of gene expression studies focus on looking for differentially expressed (DE) genes, *i.e.* genes whose mean expression is different when comparing two or more populations of samples. However, similarly to a difference of mean, a difference of variance in gene expression between sample populations may also be biologically and physiologically relevant.
RNA-sequencing (RNA-seq) has become the gold-standard technology to estimate genome-wide gene expression and the Negative Binomial distribution provides the best fit for these count data. Under this model, analyzing variance is achieved by analyzing the dispersion parameter. In the classical differential expression analysis workflow, the dispersion is only considered as a parameter to be estimated prior to looking for a difference of mean expression between conditions of interest [1]. Recently, two new methods, MDSeq [2] and DiPhiSeq [3], have been introduced to identify differences in both mean and dispersion in RNA-seq data within the same framework.

Differential expression based on RNA-seq data have been extensively studied in multiple biological contexts using different approaches these last few years. Here, we propose to evaluate MDSeq and DiPhiSeq to identify non-DE genes with a differential expression dispersion (DD) between conditions of interest. We thoroughly investigated the performances of these methods on simulated datasets [4]. In particular, we characterized the impact of some key parameters, such as the sample size per condition and the magnitude of the fold change for both mean and dispersion, on the differential dispersion detection performances between two conditions and identified settings to control the false discovery rate.

The large amount of publicly available genomic data opens new perspectives for researchers, looking for genes with a differential expression dispersion between tumor and control samples is one of those. We applied MDSeq and DiPhiSeq to The Cancer Genome Atlas (TCGA) datasets in order to identify DD and not DE mRNAs and microRNAs when comparing normal and tumor samples. Most of these genes have an increased dispersion in tumors, which may be interpreted as an overall dysregulation commonly observed in tumors. Interestingly, among DD and not DE mRNAs, the most significantly enriched Gene Ontology terms are the most widespread across all the tissues from TCGA and focus on some key cellular functions, such as catabolism. Moreover, our approach highlights some functions whose role in cancerogenesis is context-dependent, such as autophagy [5], and thus may be a lead to further investigate these biological processes.

References

[1] D. J. McCarthy, Y. Chen, and G. K. Smyth. Differential expression analysis of multifactor RNA-Seq experiments with respect to biological variation. *Nucleic Acids Res.*, 40(10):4288–4297, May 2012.

[2] D. Ran and Z. J. Daye. Gene expression variability and the analysis of large-scale RNA-seq studies with the MDSeq. *Nucleic Acids Res.*, 45(13):e127, Jul 2017.

[3] J. Li and A. T. Lamere. DiPhiSeq: Robust comparison of expression levels on RNA-Seq data with large sample sizes. *Bioinformatics*, Nov 2018.

[4] C. Soneson. compcodeR–an R package for benchmarking differential expression methods for RNA-seq data. *Bioinformatics*, 30(17):2517–2518, Sep 2014.

[5] J. M. M. Levy, C. G. Towers, and A. Thorburn. Targeting autophagy in cancer. *Nat. Rev. Cancer*, 17(9):528–542, 09 2017.

LC-MS/MS tool and interactive visualizations integration on Galaxy Workflow4Metabolomics infrastructure

Julien SAINT-VANNE[1], Romain DALLET[2], Erwan CORRE[1], Yann GUITTON[3] AND Gildas LE CORGUILLÉ[1]

[1] CNRS, Sorbonne Université, FR2424, ABiMS, Station Biologique, 29680, Roscoff, France, [2] Institut Français de Bioinformatique, CNRS UMS 3601, rue John Von Neumann, 91403 Orsay, France, [3] Laberca, Oniris, INRA,Université Bretagne Loire, 44307, Nantes, France

Corresponding Author: lecorguille@sb-roscoff.fr, yann.guitton@oniris-nantes.fr

Abstract :

Metabolomics data analysis is a complex, multistep process, which is constantly evolving with the development of new analytical technologies, mathematical methods, and bioinformatics tools and databases. The Workflow4Metabolomics [1,2] Galaxy [3] online infrastructure (W4M, https://workflow4metabolomics.org/) provides a unique centralized, user-friendly, and high-performance environment to build, run, and share metabolomics workflows for LC-MS, GC-MS and NMR technologies.

One of the major issue of the metabolomic approach is the compounds identification. To facilitate this annotation step, tandem mass spectrometry (MS/MS) is able to provide informations about the compounds structure. This technology is increasingly used by laboratories, but MS/MS data analysis remains complex and tedious. To speed up this step, various identification tools have been developed by the scientific community. For that reason, an MS/MS data processing workflow was integrated into W4M. It is based on 3 recognized tools: a data quality filtering tool, msPurity [4], and two identification tools, metFrag [5] and Sirius-CSI: FingerID [6].

During Galaxy workflows, you can access to a lot of graphic representations. All these graphical outputs were conventionally “frozen” in pdf or png format, without any real possibility for interaction. In order to make ease the results interpretations, a set of interactive visualization tools have been added to W4M tools. The recent development of Shiny applications, executable through Galaxy interactive environments, now allows interactions from graphical features and dataset filters with graphical outputs like chromatograms, heatmaps or PCA.

References :

1. Giacomoni F., Le Corguillé G. *et al.* (2014), *Workflow4Metabolomics: A collaborative research infrastructure for computational metabolomics*, Bioinformatics
2. Guitton Y. *et al.*, *Create, run, share, publish, and reference your LC–MS, FIA–MS, GC–MS, and NMR data analysis workflows with the Workflow4Metabolomics 3.0 Galaxy online infrastructure for metabolomics*, The International Journal of Biochemistry & Cell Biology, 2017, ISSN 1357-2725
3. Afgan E. *et al.*. *The Galaxy platform for accessible, reproducible and collaborative biomedical analyses: 2018 update*, Nucleic Acids Research, Volume 46, Issue W1, 2 July 2018, Pages W537–W544
4. Lawson T. *et al.*, *msPurity: Automated Evaluation of Precursor Ion Purity for MassSpectrometry-Based Fragmentation in Metabolomics*, Analytical Chemistry 2017 *89* (4), 2432-2439
5. Ruttkies, Christoph *et al.*, MetFrag relaunched: incorporating strategies beyond in silico fragmentation, *Journal of cheminformatics* vol. 8 3. 29 Jan. 2016,
6. Dührkop K., *et al.*, *Searching molecular structure databases with tandem mass spectra using CSI:FingerID*, Proc Natl Acad Sci U S A, 112(41):12580-12585, 2015

LeAFtool: Lesion Area Finding tool

Sébastien Ravel*[1,2], François Bonnot[2], and Elisabeth Fournier[2]

[1]South Green Bioinformatics Platform (SG) – Bioversity, CIRAD : UMRAGAP / BGPI / LSTM, Institut de recherche pour le développement [IRD] : UMRDIADE/ IPME – Montpellier, France
[2]Biologie et génétique des interactions plantes-parasites pour la protection intégrée – Institut national de la recherche agronomique (INRA) : UR0385, Centre de coopération internationale en recherche agronomique pour le développement [CIRAD] : UMR54 – France

Résumé

This application was created at the UMR BGPI (Unité Mixte de Recherche Biology and Genetic of Plant-Pathogen Interaction - CIRAD, INRA and Montpellier SupAgro - Montpellier, France) in order to analyse lesions on a infected leaf. An automatic measurement of lesions makes it easier to compare the pathogenicity of pathogens, whose signs and symptoms are visible. It makes possible to analyse many images in a faster way.
Scanned leaves are easily uploaded in this application. The pathogenic lesions are automatically analysed and different parameters are calculated in pixels so that the number and relative area of lesions can be calculated.

At first a code for R was developed to analyse lesion on leaves. Once it was successful, it was decided to develop an interface which could make it more accessible for all users. What makes it possible to encapsulate the code through an interface, is the R Shiny package, which allows R based web applications.
In this manual, each of the steps necessary for proper use of the interface are described. At first, the user manual provides explanations on how to install the interface. Then an explanation on how to use and manipulate the interface is given. Finally, the latest information and tips on the tool are given.

*Intervenant

Linking Allele-Specific Expression And Natural Selection In Wild Populations

Romuald Laso-Jadart*[1,2], Kevin Sugier[3], Karine Labadie[4], Emmanuelle Petit[5], Christophe Ambroise[6], Pierre Peterlongo[7], Patrick Wincker[2,8], Jean-Louis Jamet[9], and Mohammed-Amin Madoui[2,3]

[1]Génomique Métabolique, Genoscope, Institut François Jacob, CEA, CNRS, Univ Evry, Université Paris-Saclay – CEA, CNRS, Université Paris-Saclay – France

[2]Research Federation for the study of Global Ocean Systems Ecology and Evolution, FR2022/Tara Oceans GO-SEE, 3 rue Michel-Ange, 75016 Paris, France – Tara Oceans GO-SEE – France

[3]Génomique Métabolique, Genoscope, Institut François Jacob, CEA, CNRS, Univ Evry, Université Paris-Saclay – CEA, CNRS, Université Paris-Saclay – France

[4]CEA, Genoscope, Institut de Biologie François Jacob, Université Paris-Saclay, Evry, 91057, France – CEA Evry 2 rue Gaston Crémieux 91006 Evry cedex – France

[5]CEA, Genoscope, Institut de Biologie François Jacob, Université Paris-Saclay, Evry, 91057, France – Commissariat à l'Energie Atomique (CEA) – France

[6]Laboratoire de Mathématiques et Modélisation d'Evry – CNRS : UMR8071, Université d'Evry-Val d'Essonne, Institut national de la recherche agronomique (INRA) – France

[7]Univ Rennes, CNRS, Inria, IRISA - UMR 6074, F-35000 Rennes. – École normale supérieure (ENS) - Cachan, Universite de Rennes 1, CNRS : UMR6074, INRIA – France

[8]Génomique Métabolique, Genoscope, Institut François Jacob, CEA, CNRS, Univ Evry, Université Paris-Saclay – CEA, CNRS, Université Paris-Saclay – France

[9]Université de Toulon, Aix Marseille Universités, CNRS/INSU, IRD, MIO UMR 110 – CNRS : UMR110, Mediterranean Institut of Oceanography – France

Résumé

Allele-specific expression (ASE) is now a widely studied mechanism at cell, tissue and organism levels. However, population-level ASE and its evolutive impacts have still never been investigated. Here, we hypothesized a potential link between ASE and natural selection on the cosmopolitan copepod *Oithona similis*. We combined metagenomic and metatranscriptomic data from seven wild populations of the marine copepod *O. similis* sampled during the *Tara* Oceans expedition. We detected 587 single nucleotide variants (SNVs) under ASE and found a significant amount of 152 SNVs under ASE in at least one population and under selection across all the populations. This constitutes a first evidence that selection and ASE target more common loci than expected by chance, raising new questions about the nature of the evolutive links between the two mechanisms.

*Intervenant

Long-read pacbio amplicon analysis
From raw data to final results

Thomas Cokelaer[1,2], Juliana Pipoli da Fonseca[2], Anne-Sophie L'honneur[3] Flore Rozenberg[3]

[1] Institut Pasteur – Platform Biomics – 25-28 Rue du docteur Roux, Paris, France.

[2] Institut Pasteur – Bioinformatics and Biostatistics Hub – C3BI, USR 3756 IP CNRS – Paris, France

[3] Assistance Publique des Hôpitaux de Paris, Virology, Pathology and Dermatology Departments,

[4] Hôpital Cochin; Université Paris Descartes et Assistance Publique-Hôpitaux de Paris

[5] Hôpital Cochin, Service de Virologie, Institut Cochin, Inserm U1016, Université Paris Descartes, Paris, France.

Corresponding Author: thomas.cokelaer@pasteur.fr

1. Abstract

Long read technology based on Pacbio Sequel can resolve complexity of bacterial or viral populations to explore evolving genomes thanks to amplicon sequencing providing an unprecedented view of complete viral genomes. Highly accurate single molecule consensus reads gives the ability to track the evolution and phylogeny of viral populations, identify and quantify minor variants, and generate complete de novo assemblies , etc. In this work we present a flexible pipeline that was designed to analyse long-read amplicons from Pacbio Sequel technology. The pipeline addresses several problems that analysts may face when dealing with such data. First, the pipeline handles circularised data (fastq or raw data) but also raw data before circularisation. Second, it is fully parallellised and can handle tens of barcoded samples. Third, it can be tune to address several scientific questions: variant detection, phylogeny, mapping, consensus genomes. The pipeline is implemented within the Sequana library (sequana.readthedocs.io) based on the Snakemake technology (https://snakemake.readthedocs.io/en/stable/). We apply the pipeline on a set of 56 patients infected by polyomavirus. We present in this poster how the pipeline can be used on real data to extract relevant information about the viral population.

References

1. Köster, Johannes and Rahmann, Sven. “Snakemake - A scalable bioinformatics workflow engine”. Bioinformatics 2012.

2. Cokelaer et al, (2017), ‘Sequana’: a Set of Snakemake NGS pipelines, Journal of Open Source Software, 2(16), 352, JOSS http://joss.theoj.org/papers/10.21105/joss.00352

Longitudinal analysis of immune cells in kidney transplantation rejection by single-cell RNA-seq

Thomas LAURENT [1], Cynthia FOURGEUX [1] and Jeremie POSCHMANN [1]

[1] Centre de Recherche en Transplantation et Immunologie (CRTI), INSERM UMR1064, 30 Bd Jean Monnet, 44093, Nantes CEDEX 1, France

Corresponding Author: jeremie.poschmann@univ-nantes.fr

Kidney transplantation (KTx) is the most common type of transplantation surgery, yet the available donated organs do not cover the need. The number of patients on the waiting list for a KTx is continuously growing, also partly due to the graft half-life of 10 years that leads to almost 16% of patients waiting for their second transplantation. This highlights the importance of understanding the underlying processes leading to graft rejection, and more precisely chronic rejection (CR), in order to improve graft survival. Unlike acute rejections, the mechanics behind CR are poorly described, although it is the most prevalent cause of KTx failure.

In our study, we established experimental and bioinformatics tools to perform a single cell RNA sequencing (scRNA-seq) analysis, to follow gene expression within a total of 12,516 peripheral blood mononuclear cells (PBMC) across three time points in a patient with a humoral CR. The time points cover the patient condition after the KTx, during the treatment and just after the biopsy that led to the CR diagnostic. To assess the proportions of each PBMC's subtypes, a recent method of epitope labelling (CITE-Seq) [1] with barcoded antibodies was used in addition to transcriptomic information. The clusters of cells built with the shared-nearest-neighbors clustering from the Seurat [2] package are compared to the surface markers we targeted to ensure a good cell type identification. Variations of populations have been observed, like the stable increase of monocytes cells.

Moreover, we coupled it with a Cell Hashing [3] technique to multiplex samples and reduce batch effect across samples, allowing us a precise investigation of the distinct cell clusters' gene signature through time. Our findings suggest a modification of gene expression of immune cells in response to treatment following a KTx such as glucocorticoids and tacrolimus. It also highlights potential gene markers to the inflammatory response preceding a chronic graft rejection. A characterization of precursor markers for the CR may lead to a new diagnostic method with only a blood sample without the need of an invasive biopsy.

References

1. Marlon Stoeckius et al. Simultaneous epitope and transcriptome measurement in single cells. Nature Methods (14), 865, 2017.
2. Andrew Butler et al. Integrating single-cell transcriptomic data across different conditions, technologies, and species. Nature Biotechnology (36), pages 411–420,2018.
3. Marlon Stoeckius et al. Cell Hashing with barcoded antibodies enables multiplexing and doublet detection for single cell genomics. Genome Biology (19), 224, 2018.

Mechanism of mechanosensation mediated by the angiotensin II receptor 1: a molecular dynamics approach

Rym Ben Boubaker*[1], Asma Tiss[1,2], Hajer Guissouma[2], Daniel Henrion[1], and Marie Chabbert[1]

[1]Physiopathologie Cardiovasculaire et Mitochondriale – Université d'Angers, Institut National de la Santé et de la Recherche Médicale : UMR1083, Centre National de la Recherche Scientifique : UMR6015 – France

[2]Laboratoire GIPH, Faculté des Sciences de Tunis, Université de Tunis El Manar – Tunisie

Résumé

The angiotensin II receptor 1 (AT1) belongs to the superfamily of G-protein coupled receptors (GPCRs). AT1 signaling mediates the major physiological effects of angiotensin II including vasoconstriction, cardiac contractility and hypertrophy. AT1 is an important effector that controls blood pressure in the cardiovascular system. Antagonists of AT1 are broadly used for the treatment of hypertension, diabetic nephropathy and congestive heart failure. This receptor is a putative mechanosensor that can be activated by various mechanical stimuli, albeit the signaling pathways are controversial. The mechanisms leading to the activation of AT1 in response to mechanical stress are not understood and need further investigations.
In order to investigate mechanical stresses that can be sensed by AT1, we carried out molecular dynamics simulations of the receptor embedded within a hydrated POPC bilayer under NPgT conditions, using the molecular dynamics simulation program NAMD. Positive and negative values of the surface tension g led to membrane stretching and compression, respectively. We investigated the effect of both membrane stretching and compression on lipid and receptor properties.

Mechanical stress affected the physico-chemical properties of POPC, including surface area, membrane Thickness and order parameters of the lipid aliphatic chain. Stretching did not significantly alter receptor properties in the presence of allosteric sodium. However, in the absence of sodium, we could observe a transition towards a pre-activated state upon application of a surface tension. By contrast, increased pressure altered sodium binding mode and favored interaction with Asn7.49 of the NPXXY motif.
This study indicates that the molecular properties and responses of the AT1 receptor may depend on the mechanical stress present in its environment and be altered upon hypertension.

*Intervenant

Metachick: assembly and analysis of a chicken gut microbiome reveals wide variations of the caecal microbiome according to the production methods

Marie Jeammet[1], Jordi Estellé[2], Nicolas Pons[1], MetaChick Consortium and Fanny Calenge[2]
[1] MetaGenoPolis, Domaine de Vilvert, 78325, Jouy-en-Josas, France
[2] GABI, Domaine de Vilvert, 78325, Jouy-en-Josas, France

Corresponding author: marie.jeammet@inra.fr

Abstract

Between its egg and meat sectors, respectively representing an estimated 6.4 billions laying hens and 16 billions eaten birds per year, poultry production is one of the most proeminent food industries in the world - and still growing. Per extension, as the most-widely used bird for both sectors, chicken production is a major interest of poultry industrials and scientific alike, with questions ranging from growth study, disease regulation and impact of productions methods.

To answer these questions and as a result of recent discoveries regarding the insights microbiome studies can provide on their hosts, scientists have turned to metagenomic to investigate poultry microbiomes, starting with the gut, in a similar manner to what was achieved for human gut microbiome studies. Until now, 16S amplicon analysis was the most common approach, mostly because of its affordability but also due to the lack of reference genes catalog required for shotgun metagenomics and never published until now: in November 2018, the first chicken gut metagenomic catalog was released by Huang et al., exploring the five compartment of chicken digestive system and focusing on Chinese animals and on the impact of additives on their microbiotas [1].

In France, production methods are very diverse and we wanted to explore this diversity, to assemble a catalog on our own. 340 caecal samples from 34 French farms of very different production methods were thus collected and sequenced at a minimal sequencing depth of 50M reads. After relevant reads filtering and the development and adaptation of our in-house pipeline to Metachick's data type and dimensions, we then proceeded to assemble this data into a catalog of 9.7 millions non redundant genes, saturated for our samples and comparable in size and quality to the latest human catalog [2].

We then proceeded to statistically explore metagenomic samples towards a very comprehensive zootechnical data collection. Similarly to studies carried out on human and pig microbiomes, and even in taking into account confusion factors such as production sector, age seems to be the most discriminant parameter to separate individuals, young animals (mostly belonging to the meat sector) having very different profiles than older animals. Apart from age and as expected, production methods were the second most stratifying variable, with wide variations in richness, taxonomy and presence of antibiotic resistance genes, even among animals not having received antibiotics. Upon publication of the aforementioned Chinese catalog, a thorough comparison of both catalogs revealed that our catalog is more representative of the cecal metagenomic samples found on public databases than the Chinese catalog is.

Acknowledgements

This work was supported by the MetaChick Consortium. Samples collected by the Itavi. Sequencing performed by the Genoscope.

References

[1] Peng Huang. The chicken gut metagenome and the modulatory effects of plant-derived benzylisoquinoline alkaloids. *Microbiome*, 6(1):211, Nov 2018.

[2] Junhua Li and MetaHIT Consortium. An integrated catalog of reference genes in the human gut microbiome. 32(8):834–841.

Metagenomic analysis of an African beer ecosystem using FoodMicrobiomeTransfert application

Anne-Laure ABRAHAM[1], Sandra DÉROZIER[1], Quentin CAVAILLÉ[2], Thibaut GUIRIMAND[2], Solange AKA[3], Valentin LOUX[1] and Pierre RENAULT[2]

[1] MaIAGE, INRA, Université Paris-Saclay, 78350 Jouy-en-Josas, France

[2] MICALIS, INRA AgroParisTech, Université Paris-Saclay, Domaine de Vilvert, 78350, Jouy-en-Josas, France

[3] Université Nangui Abrogoua, UFR des Sciences et Technologies des Aliments, Laboratoire de Biotechnologie et Microbiologie des Aliments, 02 BP 801 Abidjan 02, Côte d'Ivoire

Corresponding Author: anne-laure.abraham@inra.fr

Tchapalo is a traditional beer produced in Ivory Coast. Its production results from a two-step fermentation of sorghum: first a spontaneous lactic fermentation yielding a sour wourt, and then, an alcoholic fermentation leading to Tchapalo. This cloudy beer has a low alcohol-content, a short shelf life (about 3 days) and its quality varies from a production to another. The precise composition of Tchapalo ecosystem is unknown, and a metagenomic approach could help to better characterize this flora and identify precisely the strains involved in Tchapalo manufacturing.

To analyze this ecosystem, we used FoodMicrobiomeTransfert, a tool we developed for metagenomic analysis of food ecosystems (http://migale.jouy.inra.fr/foodMicrobiome/). This tool, based on a mapping of the metagenomic reads on a reference genome database, identifies, for each reference genome, which genes are present in the ecosystem and gives the percentage of differences with the reference genome. It allows the user to analyze metagenomic samples via a user-friendly web interface. The user can upload metagenomes and reference genomes, choose reference genomes used for the analyze, analyze results, and share data with colleagues. Computations are performed transparently for the user on Migale platform's calculation cluster via the Bioblend API and a Galaxy portal. The web interface was developed using the Python Django framework and JavaScript for web interfaces. All the data are stored in a PostgreSQL relational database.

To illustrate the power of this tool, we will present a detailed analysis of Tchapalo ecosystem composition combining cultural, metabarcoding and metagenomic approach. In particular we performed 23 metabarcoding analysis on samples collected in different traditional producers in Abidjan, and that were further analyzed by metagenomics using ~10 million 150bp HiSeq reads. Our analysis showed that Tchapalo lactic fermentation is carried out mainly by 2 Lactobacillus species, with 5-8 other species of lactic acid bacteria present at low level. Interestingly, one of the major species appears to be poorly cultivable and its genomes reduced compared to other strains of this species. Analysis at the nucleotide level revealed that several strains of this species are present in each sample.

MetagWGS: an automated Nextflow pipeline for metagenome

Joanna FOURQUET[1], Adeline CHAUBET[2], Hélène CHIAPELLO[3], Christine GASPIN[1], Marisa HAENNI[4], Christophe KLOPP[1], Agnese LUPO[4], Jean MAINGUY[1], Céline NOIROT[1], Tony ROCHEGUE[4], Matthias ZYTNICKI[5], Tristan FERRY[6], Claire HOEDE[1]

[1] MIAT, PF Bioinfo GenoToul, Université de Toulouse, INRA, 24 Chemin de Borde Rouge, 31320 Castanet-Tolosan, France
2 INRA, US 1426, GeT-PlaGe, Genotoul, 31320 Castanet-Tolosan, France
3 MaIAGE, INRA, Université Paris-Saclay, 78350 Jouy-en-Josas, France
4 Université de Lyon, ANSES-Laboratoire de Lyon, Unité Antibiorésistance et Virulence Bactériennes, 69007 Lyon, France
5 MIAT, Université de Toulouse, INRA, 31320 Castanet-Tolosan, France
6 Hospices Civils de Lyon, Université Claude Bernard Lyon 1, International Centre for Research in Infectiology, CIRI, INSERM U1111, CNRS UMR5308, ENS de Lyon, 69364 Lyon, France

Corresponding Author: claire.hoede@inra.fr

Last decade human gut microbiota exploration allowed to identify millions of bacterial genes [1], including antibiotic resistant genes (ARGs) [2]. After antibiotic treatment, gut microbiota composition dramatically changes. Certain bacteria can exchange ARGs, spreading resistance and favoring the development of multidrug resistance [3], which is a concern for public health. Research on these complex biological DNA mechanisms is crucial.

In this project, we use whole genome shotgun sequencing to study fecal samples from gut microbiota of a healthy volunteer treated with the combination of two antibiotics for 3 days: rifampicin and levofloxacin. These pilot samples has been collected at three critical time points, before drugs administration, 3 days after the end of the antibiotic therapy and 10 days later, aiming at analysing eventual antimicrobial resistance development during antibiotic therapy. We're developing a scalable and reproducible metagenomic workflow with Nextflow [4]. It includes a Cutadapt and Sickle preprocessing step on Illumina reads, cleaning adapters and low quality reads. Cleaned reads mapping to the human genome are then removed. Quality assessment of each previous step is performed with FastQC. In addition, we compared two reads taxonomic classification programs: the well-known Kraken and a more recent tool called Kaiju [5]. We add Kaiju to the pipeline because it classifies on average 2.5 times more reads than Kraken in our data. The assembly step is performed by metaSPAdes or megahit to generate per sample contigs. Those contigs are annotated by Prokka. Then with CD-HIT we remove redundancy and generate a genes catalog by clustering ORFs at sample level and globally with a 95% sequence identity cutoff. We use BWA MEM to map reads back to contigs and featureCounts to count reads overlapping annotated genes. The raw count table gathers the number of reads aligned on each gene for each sample. A single result report is finally generated with MultiQC. We plan to improve our pipeline by adding: contig taxonomic affiliation and contig binning, ARGs and mobilome gene annotation as well as differential SNP analysis.

In conclusion, we describe a new automated metagenomic Nextflow pipeline, soon available (https://forgemia.inra.fr/genotoul-bioinfo/metagwgs) with a singularity image to ensure reproductibility and ease of use. We plan to use this pipeline in patients treated for a staphylococcal bone and joint infection with these two antibiotics, to evaluate the dynamic spread of resistance in their gut microbiota.

Acknowledgements

This project is supported by LabEx ECOFECT, Université de Lyon.

References

[1] J. Qin et al. A human gut microbial gene catalogue established by metagenomic sequencing. *Nature*, 464(7285):59–65, 2010.
[2] Y. Hu et al. Metagenome-wide analysis of antibiotic resistance genes in a large cohort of human gut microbiota. *Nat Commun.*, 4:2151, 2013.
[3] H. Nikaido. Multidrug Resistance in Bacteria. *Annu. Rev. Biochem.*, 78: 119–146, 2010.
[4] P. Di Tommaso et al. Nextflow enables reproducible computational workflows. *Nat Biotechnol.*, 35:316-319, 2017.
[5] P. Menzel et al. Fast and sensitive taxonomic classification for metagenomics with Kaiju. *Nat Commun.*, 7:11257, 2016.

Metavisitor-2, a suite of Galaxy tools for simple and rapid detection and discovery of viruses in Deep Sequence Data

Cedric MENDOZA [1], Léa BELLENGER [1], Naïra NAOUAR [1], Christophe ANTONIEWSKI[1]

[1] Plateforme de bioinformatique ARTbio, Institut de Biologie Paris Seine, CNRS, Sorbonne-Université, Inserm, 9 Quai St Bernard,75005 Paris, France

Corresponding Author: cedric.mendoza@upmc.fr

Metavisitor [1] is a Galaxy [2] tool-suite and a set of workflows for virus detection, diagnosis and discovery in Next Generation Sequencing data. The graphical Galaxy workflow editor allows users with minimal computational skills to use existing Metavisitor workflows or adapt them to suit specific needs by adding or modifying analysis modules. Metavisitor works with DNA, RNA or small RNA sequencing data over a range of read lengths and can use a combination of de novo and guided approaches to assemble genomes from sequencing reads.

In its core, Metavisitor uses *vir1* [3], a reference database of viruses we created by retrieving all viral sequences from NCBI nuccore and protein databases [4] (oct 2015). The second version of Metavisitor offers updated viruses reference database and workflows, and new tools in the Galaxy Suite.

vir2 [5] was upgraded by clustering similar and redundant sequences using the V-Clust [6] algorithm from NCBI latest release (2018). *vir2* has now 7 times less sequences than *vir1*, enabling faster alignments and viral detection with marginal loss of accuracy.

small_rna_maps [7] is a new plotting tool for small RNA alignments that allows the visualisation of read alignments to genes as well as their statistical metrics. *small_rna_maps* returns a single trellis image with all analysed samples for each gene. Observation of a large trellis of plots instead of separate images for each gene, greatly facilitates the detections of changes and differences.

Finally, a Metavisitor flavor of our deployment software GalaxyKickstart [8] is proposed, making the Metavisitor-2 tool-suite and workflows readily available for usage and easy to install, be it on a local computer or in Cloud setting.

Thus, Metavisitor-2, with its upgraded viral database *vir2*, more efficient tools and workflows and easier installation offers a better usage experience to biologists.

References

[1] Guillaume Carissimo, Marius van den Beek, Kenneth D. Vernick, and Christophe Antoniewski. 2017. Metavisitor, a Suite of Galaxy Tools for Simple and Rapid Detection and Discovery of Viruses in Deep Sequence Data. *PloS One* 12, 1 (2017), e0168397. DOI:https://doi.org/10.1371/journal.pone.0168397

[2] Enis Afgan, *et al*. 2018. The Galaxy platform for accessible, reproducible and collaborative biomedical analyses: 2018 update. *Nucleic Acids Res.* 46, W1 (July 2018), W537–W544. DOI:https://doi.org/10.1093/nar/gky379

[3] Guillaume Carissimo, Marius van den Beek, Juliana Pegoraro, Kenneth Vernick, and Christophe Antoniewski. 2016. vir1_NCBI_19-10-2015. DOI:https://doi.org/10.6084/m9.figshare.3179026.v1

[4] Dennis A. Benson, Mark Cavanaugh, Karen Clark, Ilene Karsch-Mizrachi, David J. Lipman, James Ostell, and Eric W. Sayers. 2013. GenBank. *Nucleic Acids Res.* 41, Database issue (January 2013), D36–D42. DOI:https://doi.org/10.1093/nar/gks1195

[5] Christophe Antoniewski and Cedric Mendoza. 2018. vir2_NCBI_21-03-2018. DOI:https://doi.org/10.6084/m9.figshare.6106892.v1

[6] Torbjørn Rognes, Tomáš Flouri, Ben Nichols, Christopher Quince, and Frédéric Mahé. 2016. VSEARCH: a versatile open source tool for metagenomics. *PeerJ* 4, (October 2016). DOI:https://doi.org/10.7717/peerj.2584

[7] 2019. *Collection of galaxy tools developed by the artbio-platform at the IBPS (Institut de Biologie Paris-Seine): ARTbio/tools-artbio*. ARTbio Bioinformatics facility at the IBPS. Retrieved March 28, 2019 from https://github.com/ARTbio/tools-artbio

[8] 2019. *Ansible playbooks for Galaxy Server deployment. Contribute to ARTbio/GalaxyKickStart development by creating an account on GitHub*. ARTbio Bioinformatics facility at the IBPS. Retrieved March 28, 2019 from https://github.com/ARTbio/GalaxyKickStart

METdb: A GENOMIC REFERENCE DATABASE FOR MARINE SPECIES

Guita NIANG[1], Mark HOEBEKE[1], Arnaud MENG[2], Xi LIU[1], Maxim SCHEREMETJEW[3], Rob FINN[3], Eric PELLETIER[2], Erwan CORRE[1]

[1] CNRS - Sorbonne Université - Plateforme ABIMS - Station Biologique de Roscoff, Place Georges Teissier, 29680, Roscoff, France

[2] Metabolic Genomics, Genoscope, Institut de Biologie François Jacob, CEA, CNRS, Univ Evry, Université Paris Saclay, 91000 Evry, France

[3] EMBL-EBI, Wellcome Genome Campus, Hinxton, Cambridgeshire, CB10 1SD, UK.

Corresponding Author: guita.niang@sb-roscoff.fr

The marine environment is extremely diverse and by far the largest habitat on Earth. The marine organisms are believed to be responsible for up to 98% of marine primary productivity, playing key roles in marine food webs and in carbon and energy cycles [1]. Unlike marine prokaryotes, very little genomic data from marine protists are available so far. In the context of the European project ELIXIR (https://www.elixir-europe.org/), we tends to develop (1) portable and reproducible workflows for transcriptomic assembly and annotation, (2) a transcriptomic reference database of micro-eucaryotic marine species namely METdb (http://metdb.sb-roscoff.fr/metdb/).

While part of the initial datasets originated from Marine Microbial Eukaryote Transcriptome Sequencing Project (MMETSP) [2], others come from Roscoff marine station and Tara Oceans research projects. All datasets were assembled and analyzed using the same two compartments workflow dedicated to *de novo* assembly and functional annotation, both developed with the Common Workflow Language (CWL) management system to ensure the data standardization and reproducibility. The assembly compartment includes evaluation, filtering and trimming of raw data [3] and the *de novo* assembly and evaluation of assembled transcripts. The annotation compartment defines the presence of coding regions and functional annotation is performed.

678 assemblies, generated by Lisa K. Johnson et al. [4] were recovered which corresponds to 411 marine protists (taxa). For some of these taxa, several cultivation conditions were independently sequenced, and were here co-assembled (386 datasets combined into 119 combined assemblies). These were gathered to the 292 single assemblies left and to the assemblies from 78 additional taxa from the Roscoff/Tara Ocean projects. The resource includes transcriptome assemblies and associated data (metrics and annotation) of 481 distinct marine micro-eukaryotic taxa spanning a large diversity of marine protists.

The METdb portal offers the possibility to user to explore the database by using a “simple or advanced” search function for a specific taxonomic level, a specific geographic location or a project origin and soon specific annotation. Statistical interactive charts, readsets location map and table and resulting datasets list are associated to the search functions. For each selected dataset, the user can access to both readset and assembly short summaries page with cross-references to external databases (EBI SRA, NCBI taxID, WORMS) which allows better traceability and homogeneity across databases and the possibility of downloading all resulting files.

References

[1] Kennedy J et al. Marine metagenomics: new tools for the study and exploitation of marine microbial metabolism. *Marine Drugs*, 2010.

[2] Keeling et al. The Marine Microbial Eukaryote Transcriptome Sequencing Project (MMETSP): illuminating the functional diversity of eukaryotic life in the oceans through transcriptome sequencing. *PLoS biology*,12.6, 2014.

[3] https://github.com/mscheremetjew/workflow-is-cwl/tree/assembly/

[4] Lisa K Johnson et al. Re-assembly, quality evaluation, and annotation of 678 microbial eukaryotic reference transcriptomes. *GigaScience*, 2018.

MiBiOmics, a shiny application for graph-based multi-omics analysis

Johanna Zoppi[1] and Samuel Chaffron[2,3]
[1] INSERM-UMR 1235, TENS, 1 rue Gaston Veil , 44035, Nantes, France
[2] Universite de Nantes, Centrale Nantes, CNRS-UMR 6004, LS2N, Nantes, France
[3] Research Federation (FR2022) Tara Ocean GO-SEE, Paris, France

Corresponding author: johanna.zoppi@univ-nantes.fr

Keywords

Multi-Omics, Graph-based approach, Associative Analysis, Shiny Application, Ordination.

Abstract

With the decreasing cost of sequencing, the multi-omics characterization of biological systems is becoming standard, underlying the need to design new associative methodologies. The simultaneous analysis of large heterogeneous datasets can easily become overwhelming considering the number of available techniques and the difficulty to visualize or interpret the results.Various integrative approaches have been developed in several fields including plant biology [1], microbial ecology [2], genetics [3], personalized medicine [4], and many others. Integrative multi-omics techniques are powerful approaches to create new knowledge and generate novel hypotheses. Many methods have been developed depending mostly on the experimental design, the overall research goal, and the type of data included in the analysis [5]. However these methods are usually non accessible to biologists without programming skills and do not easily provide publication-ready figures.
MiBiOmics is a Shiny web-application created to facilitate multi-omics analyses through several guided approaches, visualization tools, and an intuitive interface. The application implements classical and advanced ordination techniques as well as the reconstruction of correlation networks and their association to contextual parameters. Being complementary, these tools allow the user to infer robust results and generate new confident hypotheses.
MiBiOmics implements several ordination techniques such as PCA, PCoA, Co-inertia and Procrustes analyses and a responsive web-based version of WGCNA [6] for the easy generation and exploration of biological networks for each omics layer. In addition, we give the user the possibility to perform PLS regressions to validate the WGCNA results and further the association of omics feature to a given trait or phenotype. Key features of MiBiOmics include the dimensionality reduction of two omics dataset, the identification of common elements of interest within the networks and with respect to an external variable, and, finally, the comparison of these elements in order to find association and generate new hypotheses about the causality of a phenotype.
MiBiOmics is implemented in R, can be launched on any browser, and is freely available at https://shiny-bird.univ-nantes.fr/jzoppi/app/.

Acknowledgements

This work was supported by JOBIM 2019.

References

[1] Max Bylesjö, Daniel Eriksson, Miyako Kusano, Thomas Moritz, and Johan Trygg. Data integration in plant biology: The O2PLS method for combined modeling of transcript and metabolite data. *Plant Journal*, 52(6):1181–1191, 2007.

[2] Anna Heintz-Buschart, Patrick May, Cédric C. Laczny, et al. Integrated multi-omics of the human gut microbiome in a case study of familial type 1 diabetes. *Nature Microbiology*, 2(1):1–12, 2016.

[3] Bin Zhang, Chris Gaiteri, Liviu Gabriel Bodea, et al. Integrated systems approach identifies genetic nodes and networks in late-onset Alzheimer's disease. *Cell*, 153(3):707–720, 2013.

[4] Rui Chen, George I. Mias, Jennifer Li-Pook-Than, et al. Personal omics profiling reveals dynamic molecular and medical phenotypes. *Cell*, 148(6):1293–1307, 2012.

[5] O Paliy and V Shankar. Application of multivariate statistical techniques in microbial ecology. *Molecular Biology and Evolution*, 25(5):1032–1057, 2017.

[6] Peter Langfelder and Steve Horvath. WGCNA: An R package for weighted correlation network analysis. *BMC Bioinformatics*, 2008.

Microbial communities from deep-lake sediments of Lake Baikal, Siberia

Guillaume REBOUL[1,*], David MOREIRA[1], Paola BERTOLINO[1], Luis GALINDO[1], Natasha V. ANNENKOVA[2], Purificacion LOPEZ-GARCIA[1]

[1] Ecologie Systématique Evolution, Centre National de la Recherche Scientifique (CNRS), Université Paris-Sud, AgroParisTech, Université Paris-Saclay, Orsay, France
[2] Limnological Institute, Siberian Branch of the Russian Academy of Sciences, Irkutsk, Russia

Corresponding Author: guillaume.reboul@u-psud.fr

Lake Baikal, located in an ancient rift valley in Southern Siberia, Russia, is the deepest (maximum 1,642 meters) and largest (by volume) freshwater lake on Earth. Its surface freezes during several months in winter and water has a permanent temperature of ~4°C below the surface. Because of its depth and low temperature, Lake Baikal constitutes a privileged setting for comparison with the deep-sea sediment microbial communities. However, except for a few studies of methane hydrates and oil- and gas-emitting areas, the composition of microbial communities of deep-lake Baikal sediments remains poorly known. To fill this gap, we collected push-core sediment samples at 8 stations along a 700 km North-South transect at depths between 323 and 1450 meters. In order to have a complete overview of the entire microbial diversity and its metabolic potential, we extracted and purified the DNA before performing two different sequencing approaches. Firstly, we massively sequenced (paired-end Illumina MiSeq) 16S/18S rRNA gene amplicons covering the V4 region generated with specific prokaryotic (both archaea and bacteria) and eukaryotic primers. After cleaning low-quality sequences, we obtained 1,490,317 prokaryotic and 1,221,088 eukaryotic sequences. Clustering into OTUs (97% and 98% identity) revealed an extremely large diversity of microorganisms of the three domains of life, especially bacteria (15,384 OTUs) but also archaea (4,116 OTUs) and eukaryotes (4,196 OTUs). Secondly, we performed Illumina HiSeq Paired-end sequencing and obtained shotgun metagenomic data for each sample. Then, we conducted two different analyses on those metagenomic data: 1) we used raw reads and assembly outputs in order to compare the metagenomes in terms of microbial communities and metabolic potentials ; 2) we aimed at recovering quality MAGs (Metagenome-Assembled Genome) from the samples which have either a potential divergent position in the tree of life or interesting metabolic pathways. Despite the important difference in salinity, our study reveals commonalities with the microbial communities and metabolic pathways of deep-sea sediments, likely due to adaptation to similar physical conditions (high pressure, low temperature) and nutrients (recalcitrant organic matter for heterotrophic lineages).

MicroScope: an integrated platform for the annotation and exploration of microbial gene functions through genomic and metabolic comparative analysis

David Vallenet, Alexandra Calteau, Mathieu Dubois, Mylène Beuvin, Laura Burlot, Xavier Bussell, Stéphanie Fouteau, Guillaume Gautreau, Aurélie Lajus, Jordan Langlois, Rémi Planel, David Roche, Johan Rollin, Zoé Rouy and Claudine Médigue
Laboratoire d'Analyses Bioinformatiques pour la Génomique et le Métabolisme, Génomique Métabolique, Genoscope, Institut François Jacob, CEA, CNRS, Univ Evry, Université Paris-Saclay, 91057 Evry, France

Corresponding author: vallenet@genoscope.cns.fr

1 Introduction

Large-scale genome sequencing and the increasingly massive use of high-throughput approaches produce a vast amount of new information that completely transforms our understanding of thousands of microbial species. However, despite the development of powerful bioinformatics approaches, full interpretation of the content of these genomes remains a difficult task. To address this challenge, the LABGeM group at Genoscope has developed the MicroScope platform (https://www.genoscope.cns.fr/agc/microscope) which provides analysis for complete and ongoing genome projects together with metabolic network reconstruction and post-genomic experiments allowing users to improve the understanding of gene functions. MicroScope serves different use cases in bioinformatics:

- it supports the integration of newly sequenced or already available prokaryotic genomes through the offer of a free-of-charge service to the scientific community
- it performs computational inferences including prediction of gene function, metabolic pathways, resistome and virulome
- it provides tools for comparative genomics and metabolic analyses
- it supports collaborative expert annotation and community-based curation efforts in a rich comparative genomics context through the use of specific curation tools and graphical interfaces.

MicroScope contains data for $\sim$ 10,000 microbial genomes, which are manually curated and analyzed by microbiologists ($>$ 4,000 personal accounts in January 2019).

2 Contributions

The platform has been under continuous development since 2006 [1,2]. We will present an overview of the MicroScope analysis pipelines and illustrate the use of several new functionalities which concern:

- automatic annotation based on the UniRule system
- annotation of virulence and antimicrobial resistance genes
- comparative genomics with synteny computations and pan-genome analyses
- prediction and characterization of regions of genomic plasticity like secretion systems, integrons and secondary metabolite biosynthesis gene clusters
- metabolic network reconstruction

References

[1] Claudine Médigue, Alexandra Calteau, Stéphane Cruveiller, Mathieu Gachet, Guillaume Gautreau, Adrien Josso, Aurélie Lajus, Jordan Langlois, Hugo Pereira, Rémi Planel, David Roche, Johan Rollin, Zoe Rouy, and David Vallenet. MicroScope—an integrated resource for community expertise of gene functions and comparative analysis of microbial genomic and metabolic data. *Briefings in Bioinformatics*, sep 2017.

[2] Adrien Josso, Alexandra Calteau, Alexandre Renaux, Aurélie Lajus, David Roche, Johan Rollin, Jonathan Mercier, Mathieu Gachet, Stéphane Cruveiller, Zoe Rouy, Claudine Médigue, David Vallenet, and Claude Scarpelli. MicroScope in 2017: an expanding and evolving integrated resource for community expertise of microbial genomes. *Nucleic Acids Research*, 45(D1):D517–D528, nov 2016.

Mise en place d'un LIMS enrichi par une organisation harmonisée des métadonnées

Lysiane HAUGUEL, Fabrice DUPUIS, Sylvain GAILLARD, Julie BOURBEILLON, Claudine LANDES and Sandra PELLETIER

[1] IRHS, 42, rue Georges Morel, BP 60057, 49071 BEAUCOUZE Cedex, FRANCE

Corresponding Author: contact-bioinfo-irhs@inra.fr

À l'ère des analyses haut-débit (transcriptomique, protéomique, …), les biologistes génèrent une grande quantité d'informations qui n'est exploitée que partiellement en fonction des questions biologiques posées. Toutefois, ces données sont une source d'informations importantes non révélées par les expériences d'origine. En confrontant de grandes quantités de données issues d'expériences différentes, on peut mettre en évidence d'autres informations (études des réseaux par exemples).

Cependant une limite à cette intégration de données biologiques est l'harmonisation des métadonnées associées aux données. Pour pouvoir comparer des données entre elles, ils faut connaître certaines caractéristiques importantes pour ne pas comparer des choses radicalement différentes. La collecte de ces informations est aujourd'hui rarement automatisée et/ou structurée.

Pour permettre aux biologistes d'une unité d'organiser facilement cette collecte d'informations associées aux données, nous associons au LIMS (Laboratory Information Managment System) mis en place dans notre institut une description harmonisée de ces métadonnées afin de permettre aux outils d'intégration de données de récupérer ces informations avec un format prédéfini et une architecture organisée autour d'ontologies.

L'organisation de ces données se fait dans une base de données associée à une interface graphique.

Côté base de données, un schéma de terminologie permet de structurer les informations en fonction d'ontologies locales définies avec les biologistes sur la base d'ontologies de références.

Les ontologies utilisées pour la description d'une expérience couvrent plusieurs catégories. Nous organisons les informations liées :

- aux données sources (taxonomie, lignées, descendances,…)
- aux conditions de culture :
  - types de stress biotique et abiotiques
  - types de traitement chimique
- aux conditions de prélèvement :
  - stades de développement
  - organes / tissus

Une telle organisation nous permet ensuite de récupérer les informations associées aux données en les catégorisant suivant les ontologies choisies.

Côté interfaces graphiques, la présentations des données essaie de refléter au mieux les usages de l'utilisateur principal qu'est le biologiste.

Les données sont collectées en fonction des projets (gestion des projets, sous projets et expériences, gestion des utilisateurs et des groupes d'utilisateurs, gestion des droits en lecture et en écriture). Les échantillons sont groupés en fonction des projets et organiser en arbre à partir du prélèvement jusqu'aux résultats. Ainsi, à partir d'un projet, la visualisation des échantillons sous forme d'arborescence regroupe l'ensemble des échantillons techniques (ARN, protéines,…) issue d'un même prélèvement, et permet la sélection rapide des résultats en fonction de leurs type (transcriptome, protéome, phénotypage,...)

Mise en place d'un pipeline automatisé d'analyses multivariées pour la cytométrie en flux multi-couleurs

Alexia ALFARO, Mélanie GUYOT, Clara PANZOLINI, Philippe BLANCOU, Agnès PAQUET, Julie CAZARETH

Université Côte d'Azur, CNRS, Institut de pharmacologie moléculaire et cellulaire

alexiaalfaro@hotmail.fr,cazareth@ipmc.cnrs.fr

Introduction :

Les analyses standards de cytométrie en flux multi-couleurs sont basées sur une stratégie hiérarchique de sélection manuelle des populations de cellules en fonction des niveaux d'expression des différents marqueurs (gating). Néanmoins, il existe de nombreux outils bioinformatiques qui permettent une exploration non-supervisée des données, assurant une analyse exhaustive et reproductible. Le but de notre étude est la mise en place d'un pipeline d'analyses non-supervisées performant, et utilisable par les biologistes au sein de la plateforme de cytométrie de l'institut.

Méthode :

La mise en place du pipeline et les tests de performance ont été réalisés sur un jeu de données constitué de 60 échantillons de souris, comparant 2 conditions expérimentales sur une cinétique de 5 jours, analysés à l'aide de 14 marqueurs permettant d'identifier les cellules de l'immunité innée. Le contrôle qualité des échantillons, ainsi qu'un gating manuel des populations majoritaires connues, servant de dataset de référence, ont été réalisés avec le logiciel FlowJO. Les analyses non-supervisées ont été réalisées avec FlowSOM (Van Gassen et al 2015). Les données ont été visualisées via un tSNE réalisé sur un sous-échantillonnage des data à 5,000 cellules/échantillon. Les analyses ont été effectuées dans l'environnement R et adaptées pour les biologistes avec le package Cytofkit (Chen et al 2016).

Résultats :

Les paramètres de FlowSOM ont été optimisés en comparant les pourcentages de cellules identifiées avec FlowSOM à ceux de notre dataset de référence. L'expression des marqueurs visualisés sur le tSNE permet d'identifier les populations selon le phénotype cellulaire attendu avec le gating manuel. Nous avons également déterminé les limites de détection de populations rares en créant des jeux de données tests à partir de sous-échantillonnages de notre dataset de référence. Par exemple, pour les neutrophiles, représentant 18,53% des cellules de départ, et présents dans chaque condition expérimentale, nous avons pu détecter la population jusqu'à 0,15% de la population totale. Nous avons également testé la capacité à détecter une population rare présente dans une seule des conditions expérimentales. FlowSOM a réussi à identifier une population de monocytes circulants présente de 4,89% à 0.63% des cellules, sans attribuer ces cellules à l'autre condition ni les mélanger dans d'autres clusters.

Conclusion :

Les tests de performance de FlowSOM que nous avons effectués montrent que dans nos conditions expérimentales, cet algorithme d'analyse non-supervisée est capable de regrouper les populations cellulaires de façon similaire à ce qui est réalisé manuellement, même les populations rares. Le pipeline est maintenant opérationnel sur la plateforme de cytométrie à la fois en ligne de commande R et par Cytofkit.

MobiDL: next generation family of WDL DNA-NGS pipelines

David Baux[1], Nicolas Soirat[1], Kevin Yauy[2], Thomas Guignard[2], Henri Pégeot[1], Olivier Ardouin[1,2,3], Charles Van Goethem[3], Michel Koenig[1] and Anne-Françoise Roux[1]

[1] Laboratoire de génétique moléculaire des maladies rares, EA7402, Université de Montpellier, Laboratoire de génétique moléculaire, CHU de Montpellier, Université de Montpellier, Montpellier, France

[2] Unité de Génétique Chromosomique, Hôpital Arnaud de Villeneuve, CHU de Montpellier, Université de Montpellier, Montpellier, France

[3] Laboratoire de Biologie des Tumeurs Solides, Hôpital Arnaud de Villeneuve, CHU de Montpellier, Université de Montpellier, Montpellier, France

Corresponding author: david.baux@inserm.fr

The treatment of DNA-NGS data in production environment is becoming a crucial issue in diagnostic laboratories as more and more clinical tests involve massively parallel sequencing (MPS) methods (mainly Illumina). Many of these tests use targeted gene sequencing (gene panels), which still remains method of choice because of costs, practical and technical reasons. However, medium throughput instruments (e.g. Illumina NextSeq) can now run up to 96 samples in a single experiment for small panels. Therefore efficient data treatment must require optimized, reproducible, portable, validated, parallelized pipelines able to run in HPC environment.

The bioinformatics group of Montpellier University Hospital (MoBiDiC group) is developing a family of pipelines to help fulfill these needs.
Currently two pipelines are available for production use:

— MobiDL panelCapture for secondary analysis of gene panels
— MobiDL captainAchab for tertiary analysis

These pipelines use the WDL workflow language which in turn uses the Cromwell execution engine. The WDL language is able to easily slice data in order to parallelize non multi-threaded tools (e.g. GATK Haplotype Caller). The Cromwell engine is flexible enough to support execution on personal computers, many HPC schedulers and public clouds.

MobiDL panelCapture follows the Broad Institute best practices for variant discovery. It includes as unusual features sambamba for faster processing of indexing and duplicate marking of alignment files and also outputs them directly in CRAM format. The workflow has been validated using real and simulated data (Table 1).

Tab. 1. MobiDL panelCapture validation against simulated reads

Library size (Mb)	Mean DoC	#variants	F-measure
1	100X	584	1.0000
16	50X	16254	0.9895

Future improvements will include the integration of a second variant caller and the release of a singularity container to facilitate the portability.

The second workflow captainAchab aims at annotating, ranking and visualizing the results of a secondary analysis. It takes as input a VCF file and a list of HPO codes to prioritize the genes according to the phenotypes (using phenolyzer). CaptainAchab then runs ANNOVAR and MPA for variant annotation and ranking. Finally, the workflow uses the Captain-ACHAB script to merge all the data into a single spreadsheet file. It is able to run and prioritize trios analyses, and prioritize variants taking into account the mode of inheritance. This workflow is available as WDL scripts but also as a singularity container (achabilarity).

MobiDL workflows are powerful DNA-NGS analysis pipelines and come with companion shell scripts that facilitate their automation and deployment in a production environment. They are available on GitHub under a GNU GPL v3.0 license.

Modelling the differentiation dynamics of monocytes in contact with CLL B cells

Hélène Arduin, Marcin Domagala, Mary Poupot and Vera Pancaldi
Cancer Research Center of Toulouse UMR 1037 InsermUniversité Toulouse III Paul Sabatier,
2 avenue Hubert Curien, 31037 Toulouse Cedex 1, France

Corresponding author: helene.arduin@inserm.fr

Monocytes are immune cells which can differentiate into macrophages to help defend the body against pathogens. Macrophages are polarized along a spectrum with two extreme phenotypic states: the M1 phenotype, pro-inflammatory and stimulating the immune system, and the M2 phenotype, anti-inflammatory and stimulating tissue repair [1]. In a tumour setting, some macrophages can become strongly linked to cancer cells. These are called tumour-associated macrophages and are mostly polarized towards the M2 phenotype [2]. In the case of chronic lymphocytic leukemia (CLL), tumour-associated macrophages are called nurse-like cells (NLCs). They reside mainly in the lymph nodes, where they protect leukemic B cells (BCLL) from spontaneous apoptosis and contribute to their chemoresistance [3,4]. NLC are differentiated from monocytes through contact with BCLL and soluble factors [5], however, the precise mechanisms by which BCLL influence this differentiation are still unknown.

Here we propose an agent-based model (ABM) of monocyte differentiation in a BCLL culture. The goal is to study the conditions and dynamics of monocytes differentiation, depending on monocytes and BCLL initial relative densities and on their sensitivity to other cells' presence.

Five kinds of agents are represented in the ABM: BCLL (i.e. cancer cells), monocytes, macrophages, and NLCs. Upon initialisation, only BCLL and monocytes are present, but during the course of the simulation, monocytes will differentiate into either macrophages or NLCs. BCLL are attracted towards NLCs in order to receive an apoptosis-blocking signal. If they end up too far from any NLC they will die, so the initial density of monocytes and the parameters governing the differentiation are paramount here. That is why we are also conducting *in vitro* experiments of BCLL and monocytes co-culture, in order to establish the minimum monocytes and cancer cells densities that are necessary to ensure cell survival, and to establish whether the initial ratio between monocytes and BCLL is of importance. From these experiments, we obtain data on the final relative proportions of each cell type in the culture, after differentiation is complete and for 11 different initial density conditions. With these data and the help of model exploration techniques, we have started to test and validate our ABM.

This model is a first step to a better understanding of monocyte differentiation in a tumoral environment, and in particular of the way in which cancer cells can influence monocytes to differentiate into pro-tumoral macrophages. A more complex model of the entire tumour micro-environment, taking into account different kinds of immune cells (for instance T cells, which are usually needed for immunotherapy success) and integrating the signaling pathways that lead to monocyte differentiation, is the next step.

References

[1] Peter J. Murray. Macrophage Polarization. *Annu. Rev. Physiol.*, 79(1):541–566, 2017.

[2] Antonio Sica, Paola Larghi, Alessandra Mancino, Luca Rubino, Chiara Porta, Maria Grazia Totaro, Monica Rimoldi, Subhra Kumar Biswas, Paola Allavena, and Alberto Mantovani. Macrophage polarization in tumour progression. *Semin. Cancer Biol.*, 18(5):349–355, 2008.

[3] Jan A Burger, Nobuhiro Tsukada, Meike Burger, Nathan J Zvaifler, Marie Dell Aquila, J Thomas, Washington Dc, and Thomas J Kipps. Blood-derived nurse-like cells protect chronic lymphocytic leukemia B cells from spontaneous apoptosis through stromal cell – derived factor-1. *Blood*, 96(8):2655–2663, 2000.

[4] F Boissard, J-J Fournié, A Quillet-Mary, L Ysebaert, and M Poupot. Nurse-like cells mediate ibrutinib resistance in chronic lymphocytic leukemia patients. *Blood Cancer J.*, 5(10):e355–e355, 2015.

[5] Frédéric Boissard, Jean Jacques Fournié, Camille Laurent, Mary Poupot, and Loïc Ysebaert. Nurse like cells: Chronic lymphocytic leukemia associated macrophages. *Leuk. Lymphoma*, 56(5):1570–1572, 2015.

Molecular Modeling of the Asc-1 Transporter: Insights into the first steps of the transport mechanism

Afaf Mikou[a], Alexandre Cabayé[a,b], Anne Goupil[b], Hugues-Olivier Bertrand[b], Jean-Pierre Mothet[c], Francine Acher[a]

[a] *UMR 8601 CNRS, Université Paris Descartes*

[b] *BIOVIA, Dassault Systèmes*

[c] *UMR 9188 CNRS-ENS, Université Paris Saclay*

afaf.mikou@parisdescartes.fr

Elucidating the transport mechanism of the Alanine-Serine-Cysteine Transporter (Asc-1, SLC7A10) is central for our basic understanding of the function of this critical transmembrane protein, and for the subsequent rational design of drugs for the treatment of Schizophrenia particularly but also for other brain disorders. However, to date, the mechanism implicated in the regulation of Asc-1 and its shuttle of amino acids across cell membranes remain unclear.

The Asc-1 Transporter is the light chain of the heterodimer transporter (HAT) and is a Na^+ independent antiporter distributed throughout the CNS [1]. Its primary role is the regulation of the synaptic availability of D-Serine, which is central for brain communication by serving as the main co-agonist for NMDA receptors. Asc-1 operates as a facilitative transporter and as an antiporter and controls the extracellular levels of D-Serine in the CNS [2,3]. Although little is known about the atomic-level details of its regulation and shuttle properties, we reasoned that a better understanding of the conformational changes experienced by Asc-1 during the transport cycle would be informative for drug development.

Here, we employed homology modeling and molecular dynamics calculations to explain some of Asc-1 properties. Given that Asc-1 crystal structures are not yet available, we built homology models of the various states based on X-ray structures of similar transporters. Despite a low sequence identity (less than 20%) with other proteins in the SLC7 family, all these transporters nonetheless adopt a conserved 5+5 inverted topology fold and 3D structural homology. Based on this, the three-dimensional Asc-1 "outward-open model" (apo, substrate free) and Asc-1 D-Serine "bound occluded model" were built using sequence and structural alignment of crystal coordinates of AdiC free (PDB code 5J4I) and AdiC Arg-bound occluded (PDB code 3L1L), respectively (**Figure 1**) [4,5].

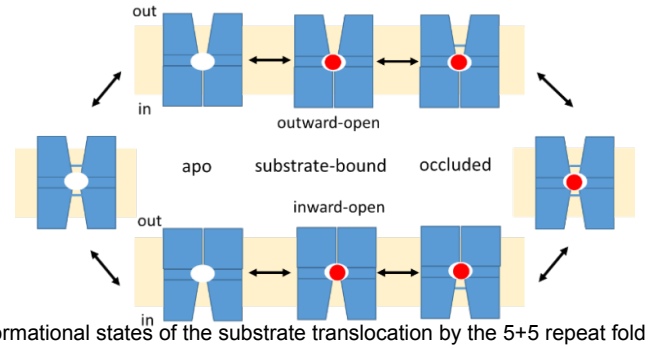


Figure 1: Conformational states of the substrate translocation by the 5+5 repeat fold transporters [6]

We then focused on the transition between Asc-1 outward-open and Asc-1 occluded conformations in complex with D-Serine to better understand the process by which this transporter carries a substrate from the periplasm to the cytoplasm across the cell membrane. To do so, we used the GOLD program [7] as implemented in Discovery Studio (BIOVIA, Dassault Systèmes) to search for the best pose for docking D-serine in the potential binding site. We then ran molecular dynamic calculations to identify key amino acids involved in the substrate translocation. Several backbone-backbone interactions were found between the protein and D-Serine substrate defining the binding site which was in good agreement with those found in the AdiC Arg-bound complex. Also, major conformational changes were observed especially in TM6 including Phe243, which are believed to guide D-serine in the course of its translocation. Interestingly, this matches quite well with that observed for the corresponding residue Trp202 in AdiC, which interacts with an aliphatic portion of the substrate to block the exit route back to the periplasm.

In conclusion, we believe that our models can help understanding transport mechanisms of Asc-1 which can then be tested experimentally and used for rational drug discovery efforts.

References

1. M. Palacin et al. Biochem Soc Tran. **(2016)**, 44, 745.
2. JP. Mothet et al. J.Neurochem. **(2015)**, 135, 210.
3. H. Sason, et al. Cereb Cortex **(2017),** 27, 1573.
4. X. Gao et al. Nature **(2010)**, 463, 828.
5. EM. Krammer et al. PloS One **(2016)**, 11.
6. L. Kowalczyk et al. PNAS **(2011)**, 108, 3955.
7. G. Jones et al. J Mol Biol. **(1997),** 267, 727.

Multi-factor Data Normalization enables the detection of LOH in amplicon sequencing data

Rim ZAAG [1], Maxime LIENARD [1], SÉBASTIEN TOFFOLI [2] and Jean-François LAES [1]

[1] OncoDNA, 1 Rue Luis Breguet, 6041, Gosselies, Belgium

[2] Institute of Pathology and Genetics, 25 Avenue Georges Lemaître, 6041, Gosselies, Belgium

Corresponding Author: r.zaag@oncodna.com

The emergence of amplicon sequencing has revolutionized cancer diagnostics and treatment. While whole exome sequencing (WES), performing in-depth sequencing of nearly all the coding exons, is relatively expensive, the amplicon sequencing technique allows sequencing a limited number of exons at a low price. Amplicon sequencing is used in routine to detect point mutations (single-nucleotide variants or small size insertions/deletions), copy number alterations (CNAs) and translocations, but is not commonly used for loss of heterozygosity (LOH) detection despite its important role in cancer. Beyond its role as prognosis factor, LOH has recently been reported to play a key role in precision medicine, as it can predict response to immune checkpoint blockade therapies [1]. The current LOH detection methods mainly rely on the comparison of the tumor genotype against its normal counterpart using SNP markers, by detecting the transition from a heterozygous locus in normal DNA to a homozygous state at that same locus in cancer cells [2,3,4,5]. Nonetheless, the use of these methods in routine diagnosis is limited by the frequent lack of paired normal DNA for some patients and the low tumor purity. Some studies were able to detect LOH by genotyping only the tumor, exploiting high-throughput SNP array techniques [6]. However, only 30% of SNPs in an individual sample are heterozygous, which makes the remaining 70% non-informative [7]. On the other hand, the identification of copy number loss (CNA-loss) has been used for the detection of LOH and can be a response but it remains incomplete, especially for neutral LOH cases and the distinction between the loss of one or both copies of a gene (copy number variations). As traditional methods of normalization hide small differences of the data, it is very difficult to detect the loss of two copies, and very often the output refers to the homozygous deletion as the loss of one copy.

Here we address this gap and we present ALHASCA (Algorithm for LoH identification in Amplicon Sequencing in CAncer), an algorithm that includes a methodology to design assays and a multifactor normalization and annotation technique enabling the detection of large LOH and copy number losses from amplicon sequencing data. ALHASCA was developed to handle unpaired tumor samples with different tumor impurity levels by (i) defining a methodology for the design of gene panels and the selection of a set of highly polymorphic informative SNP markers, (ii) defining a method to normalize read coverage at different scales to address the intra-library and the technology-specific variability and (iii) assigning statistical significance to putative LOH and/or CNVs resulting from the segmentation of normalized profiles. We have validated the proposed algorithm on a high-depth tumor-only sequencing data for 10 samples for which array CGH profiles were available. We showed that the results obtained from the ALHASCA method compare favorably with gold standard by accurately detecting all expected LOH events with high specificity and precision of 86% and 80% respectively.

References

1. Sade-Feldman, Moshe and Jiao, Yunxin J. and Chen, Jonathan H. and Rooney, Michael S. and Barzily-Rokni, Michal and Eliane, Jean-Pierre and Bjorgaard, Stacey L. and Hammond, Marc R. and Vitzthum, Hans and Blackmon, Shauna M. and Frederick, Dennie T. and Hazar-Rethinam, Mehlika and Nadres, Brandon A. and Van Seventer, Emily E. and Shukla, Sachet A. and Yizhak, Keren and Ray, John P. and Rosebrock, Daniel and Livitz, Dimitri and Adalsteinsson, Viktor and Getz, Gad and Duncan, Lyn M. and Li, Bo and Corcoran, Ryan B. and Lawrence, Donald P. and Stemmer-Rachamimov, Anat and Boland, Genevieve M. and Landau, Dan A. and Flaherty, Keith T. and Sullivan, Ryan J. and Hacohen, Nir. Resistance to checkpoint blockade therapy through inactivation of antigen presentation. Nature Communications, 1136, 2017.
2. *Lin, Ming C., Lee-Jen Wei, William R. Sellers, Marshall Lieberfarb, Wing Hung Wong and Cheng Li. dChipSNP: significance curve and clustering of SNP-array-based loss-of-heterozygosity data. Bioinformatics 20 8, 1233-40, 2004.*
3. *Koed, Karen, Carsten Wiuf, Lisbeth Bredholt Christensen, Friedrik Wikman, Karsten Zieger, Klaus Meyer Møller, Hans von der Maase and Torben Falck Orntoft. High-density single nucleotide polymorphism array defines novel stage and location-dependent allelic imbalances in human bladder tumors. Cancer research 65 1, 34-45, 2005.*
4. *Lamy, Philippe, Claus Lindbjerg Andersen, Lars Dyrskjot, Niels Torring and Carsten Wiuf. A Hidden Markov Model to estimate population mixture and allelic copy-numbers in cancers using Affymetrix SNP arrays. BMC Bioinformatics 8, 434 – 434, 2007.*
5. *Rancoita, Paola M. V., Marcus Hutter, Francesco Bertoni and Ivo Kwee. An integrated Bayesian analysis of LOH and copy number data. BMC Bioinformatics,* 11(1):321, *2010.*
6. *Beroukhim, Rameen, Mengshi Lin, Yuhyun Park, Ke Hao, Xiaojun Zhao, Levi A. Garraway, Edward Alan Fox, Ephraim P. Hochberg, Ingo K. Mellinghoff, Matthias D. Hofer, Aurelien Descazeaud, Mark A. Rubin, Matthew Meyerson, Wing Hung Wong, William R. Sellers and Cheng Li. Inferring Loss-of-Heterozygosity from Unpaired Tumors Using High-Density Oligonucleotide SNP Arrays. PLoS Computational Biology 2, 820 – 823, 2006.*
7. *Wu, Ling-Yun, Xiaobo Zhou, Fuhai Li, Xiaorong Yang, Chung-Che Jeff Chang and Stephen T. C. Wong. Conditional random pattern algorithm for LOH inference and segmentation. Bioinformatics 25 1, 61-7, 2009.*

Multi-omics approach to predict drug response in liver cancer cell lines

Stefano CARUSO*[1,2], Anna-Line CALATAYUD*[1,2], Jill PILET*[1,2], Samia REKIK[1,2], Tiziana LA BELLA[1,2], Sandrine IMBEAUD[1,2], Eric LETOUZÉ[1,2], Léa MEUNIER[1,2], Quentin BAYARD[1,2], Nataliya ROHR-UDILOVA[3], Camille PÉNEAU[1,2], Bettina GRASL-KRAUPP[4], Leanne DE KONING[5], Bérengère OUINE[5], Paulette BIOULAC-SAGE[6,7], Gabrielle COUCHY[1,2], Julien CALDERARO[8], Jean-Charles NAULT[1,2,9], Jessica ZUCMAN-ROSSI[§1,2,10#], Sandra REBOUISSOU[§1,2#.]

[1]Centre de Recherche des Cordeliers, Sorbonne Universités, Inserm, UMRS-1138, F-75006 Paris, France
[2]Functional Genomics of Solid Tumors, USPC, Université Paris Descartes, Université Paris Diderot, Université Paris 13, Labex Immuno-Oncology, équipe labellisée Ligue Contre le Cancer, F-75000 Paris, France
[3]Division of Gastroenterology and Hepatology, Clinic of Internal, Medicine III, Medical University of Vienna, Vienna, Austria
[4]Department of Medicine I, Division: Institute of Cancer Research, Comprehensive Cancer Center, Medical University of Vienna, Vienna, Austria
[5]RPPA platform, Curie Institute, PSL research university, Paris, France;
[6]Bariton INSERM, UMR-1053, Bordeaux, France
[7]Department of Pathology, Pellegrin Hospital, Hospital of Bordeaux, Bordeaux, France
[8]Anathomopathology Department, Henri Mondor Hospital, Créteil; University of Paris Est Créteil, Inserm U955, Team 18, Mondor Institute of Biomedical Research, France
[9]Liver unit, Jean Verdier Hospital, University Hospitals Paris-Seine-Saint-Denis, AP-HP, Bondy, France
[10]European Hospital Georges Pompidou, AP-HP, F-75015, Paris, France

*[§] These authors contributed equally to this work

Corresponding Author: stefano.caruso@inserm.fr | sandra.rebouissou@inserm.fr | jessica.zucman-rossi@inserm.fr

Abstract: Hepatocellular carcinoma (HCC) is a heterogeneous and aggressive malignancy with poor prognosis at advanced stage. Due to this molecular heterogeneity, the current therapies have low efficacy and limited survival benefit. The aim of this study was to use a multi-omics approach in a large panel of liver cancer cell lines in order to capture the molecular diversity of HCC and predict the drug sensitivity.

We performed whole-exome, RNA and microRNA sequencing and quantification of 126 proteins across 34 liver cancer cell lines screened with 31 anti-cancer compounds. Correlation analysis and Elastic net regression were used to identify molecular features associated with drug sensitivity. Molecular profiles of liver cancer cell lines and HCC primary tumors were compared.

The 34 liver cancer cell lines harbored the common genetic alterations identified in the more proliferative and aggressive HCCs. Unsupervised consensus classification identified three robust transcriptomic subgroups related to the differentiation status and associated with the diversity of therapeutic responses, with the most differentiated CL1 subgroup showing the highest drug sensitivity. Elastic net regression yielded a huge number of molecular markers related to drug response with a median of 95 associated features per drug [0-139] and uncovered strong associations. In particular, we found the expression of 5 genes (HSD17B7, RORC, MRPS14, SERINC2, LAD1) that predict accurately the response to the MEK1/2 inhibitors trametinib and refametinib.

Correlation analysis identified specific drugs that could target HCCs with distinct molecular features such as inactivating mutations in TSC1/TSC2 and TP53 associated with higher sensitivity to the mTOR inhibitor rapamycin and the AURKA inhibitor alisertib.

This study provides a comprehensive molecular characterization of the most widely used liver cancer cell lines and identify specific molecular features associated with distinct drug responses that may be useful to stratify patients in future clinical trials. In addition, all the data are freely accessible to the scientific community (www.zucmanlab.com, available from July 2019).

Acknowledgements: We thank Giuseppe Maltese and Codebase for the valuable support during the website building process.

MYCMACS - MYCétes pour une Meilleure Acquisition des Connaissances Scientifiques

Alfred GOUMOU[1], Coralie ROHMER[1], Marie GRISON[1], Marie-Anne LE MOIGNE[1], Thomas GUILLEMETTE[1], Valérie GRIMAULT[2], Claudine LANDES[1] and Sylvain GAILLARD[1]

[1] IRHS, 42 avenue Georges Morel, 49070, Beaucouzé, France

[2] GEVES, 25 avenue Georges Morel, 49070, Beaucouzé, France

Corresponding Author: sylvain.gaillard@inra.fr

1. Contexte

MYC-MACS est un projet à la fois à visée pédagogique et de recherche basé sur l'étude des champignons de type micromycète. Certains micromycètes peuvent particulièrement être utiles pour l'homme. Il est important également de pouvoir caractériser ceux qui sont nuisibles. Ce projet est développé conjointement par des étudiants de biologie, des enseignants-chercheurs de l'Université d'Angers et des chercheurs de l'INRA, en collaboration avec les plateaux techniques de la Structure Fédérative de Recherche « Qualité et Santé du Végétal », le GEVES et Terre des Sciences. L'objectif principal est d'explorer la diversité fongique issue d'environnements extrêmement variés afin de générer et partager à un niveau local, national et international du matériel éducatif et scientifique. Les mycètes sont d'abord isolés à partir d'échantillons biologiques variés puis identifiés par des critères morphologiques et moléculaires pour un positionnement phylogénique précis. Nous nous focalisons en priorité sur les mycètes associés aux plantes. Chaque spécimen est photographié permettant l'acquisition d'une large collection d'images réalisées en microscopie optique et électronique. L'ensemble de ces données aboutira à créer une base de données évolutive accessible en français et en anglais via des médias numériques et constituant un outil pédagogique innovant et original destiné à la connaissance et l'identification des mycètes. Les spécimens biologiques isolés sont par ailleurs conservés au sein d'une collection et mis à disposition de la communauté scientifique et éducative. Les images, cultures fongiques et séquences génomiques serviront de support pour des actions de diffusion du savoir scientifique vers le grand public et les lycées ainsi que pour des finalités de recherche (diagnostic moléculaire par exemple).

2. Réalisation

Nous présenterons un prototype d'outil développé dans le cadre du projet MYC-MACS qui permet via une application Android d'identifier un micromycète à partir d'une clé d'identification simplifiée. Les différentes étapes d'identification sont illustrées par des photographies explicitant les clés : couleur de la colonie sur boite de Petri, forme des spores par exemple, pour aider à la décision.

L'application est développée en Javascript en exploitant le framework Qooxdoo mobile. La base de données est embarquée sous format JSON afin d'être disponible sur le terrain et en salle de TP dans les lycées ou les universités dans des environnements pas nécessairement accessibles en WiFi tout en permettant sa mise à jour simple par téléchargement d'un fichier. Cette application pourra également être utilisée dans des manifestations à visée large public comme la Fête de la science ou la Journée des chercheurs organisées par l'association Terre des Sciences.

Les isolats ont été obtenus dans le cadre de TP collaboratifs par des étudiants de l'Université d'Angers. Les photographies en microscopiques optiques ont été obtenues sur le plateau IMAC de la SFR QUASAV. Le GEVES met à disposition sa collection de micromycètes ainsi que le plateau COMIC. La clé d'identification simplifiée est développée par l'équipe FungiSem de l'IRHS, spécialisée dans l'étude des champignons associés aux semences. Des fiches descriptives des espèces identifiées sont fournies comme support pédagogique avec une visée grand public. Ces fiches sont également rédigées dans le cadre de TP collaboratifs.

Acknowledgements

Ce projet est soutenu par la région Pays de la Loire qui finance les stages de M2 de C. Rohmer et A. Goumou.

mzLabelEditor: un outil pour annoter des spectres de masse

Virginie LOLLIER[1], Mathieu FANUEL[1], Dominique TESSIER[1] and David ROPARTZ[1]

[1] INRA UR1268 Biopolymères Interactions Assemblages, F-44316 Nantes, France

Corresponding Author: virginie.lollier@inra.fr

Dans le domaine de la glycomique, la résolution des structures de glycanes par spectrométrie de masse fait l'objet de développements méthodologiques et repose essentiellement sur l'expertise des scientifiques. Il n'existe pas d'outil dédié qui faciliterait l'analyse des spectres de masse et l'annotation se fait essentiellement en mêlant des captures d'écran et des zones de texte dans un outil pour la bureautique. Pour faciliter cette annotation et capitaliser sur les connaissances accumulées au fil des développements méthodologiques, nous avons développé une application graphique pour l'annotation manuelle mais standardisée des spectres de masse.

MzLabelEditor permet l'affichage d'un spectre sur lequel deux types d'étiquettes, les indications de masses et les intitulés des ions, sont accrochées à des pics de masse (m/z). L'utilisateur peut éditer chaque annotation et la déplacer sur le spectre pour la rendre plus lisible. Des filtres sont disponibles pour afficher ou masquer des lots d'annotations. L'utilisateur peut également intervenir sur les caractéristiques d'affichage du spectre, comme la taille des échelles aux abscisses et aux ordonnées, sans perdre les annotations existantes.

L'application mzLabelEditor est généralisable aux annotations de tous types de données MS/MS indépendamment de l'appareil utilisé pour l'acquisition, du constructeur mais également du type de molécules. Le spectre traité est notamment associé à un formulaire de description des conditions instrumentales mises en œuvre. La description du type d'appareil utilisé, du mode de fragmentation et d'ionisation des molécules s'appuie sur l'ontologie pour la spectrométrie de masse développée par l'HUPO-PSI (Human Proteome Organization-Proteomics Standards Initiative [1]). L'application utilise le format texte pour la lecture et l'écriture des informations de manière à ce qu'elles restent lisibles et réutilisables dans n'importe quel éditeur de texte. Elle peut également parcourir les données de spectrométrie de masse à partir des formats basés sur XML tels que mzML ou mzXML. Pour permettre la publication des annotations sous forme d'images de bonne qualité, l'export du spectre au format svg est proposé.

L'application, développée en Java8, utilise JavaFX pour l'interface graphique et des librairies externes telles que JFreeChart, ApacheJena et jmzreader [2], pour la manipulation des données.

References

[1] Gerhard Mayer, Luisa Montecchi-Palazzi, David Ovelleiro *et al.*. The HUPO proteomics standards initiative- mass spectrometry controlled vocabulary. *Database*, bat009, 2013.

[2] Johannes Griss, Florian Reisinger, Henning Hermjakob, Juan Antonio Vizcaíno. jmzReader: A Java parser library to process and visualize multiple text and XML-based mass spectrometry data formats. *Proteomics*, 12(6):795-8, 2012.

Navigating the treacherous waters of HLA imputation with the SHLARC (SNP-HLA Reference Consortium)

Venceslas DOUILLARD[1,2], Nicolas VINCE[1,2], Sophie LIMOU[1,2,3] and Pierre-Antoine GOURRAUD[1,2]

[1] ATIP-Avenir, Centre de Recherche en Transplantation et Immunologie UMR 1064, INSERM, Université de Nantes, Nantes, France

[2] Institut de Transplantation Urologie Néphrologie (ITUN), CHU Nantes, Nantes, France

[3] Ecole Centrale de Nantes, Nantes, France

Corresponding Author: nicolas.vince@univ-nantes.fr

The HLA region is crucial in the understanding of a lot of pathologies as suggested by the high number of associations with immune-related diseases. Although SNPs association studies grew importantly in the last decade, direct HLA allele association is hindered by the complexity of typing. HLA imputation offers a statistical alternative to current HLA typing, cutting costs and time alike. The power of this method relies on machine learning models obtained with the R package HIBAG[1] which are generated from individuals with known SNP+HLA data and allows prediction of HLA alleles from SNPs. The composition of these models (or reference panels) is crucial to achieve high imputation accuracy as different populations differ in both SNP and allele diversity and/or frequency.

The aim of SHLARC is to gather immunogeneticists on a platform to build and share large public reference panels with anonymized data or directly impute HLA from SNPs, using our expertise and access to supercalculator. Indeed, our current work on reference panels highlighted the effect of: 1) number of individuals, we showed a two-fold increase in accuracy (average of 45% to 86% with 10 to 100 individuals, respectively); 2) number of SNPs, accuracy went from 75% to 86% with 100 to 5,000 SNPs, respectively. Additionally, we could reduce the adequate number of SNP in the model by 1.5 to 25 times by creating a custom reference panel where only a specific subset of SNP available in the data to impute were selected. This reduced greatly the computation time and limited the need to prior SNP imputation.

We will pursue our effort to assess the relevance of population matching before modelling and uncover specific SNPs which may be essential for imputation. Our access to 1000 Genomes data as well as an African-American population gave us a better grasp of HLA imputation, however we truly believe sharing SNP+HLA data in a global consortium will be beneficial to all and place HLA association at the forefront of immunogenomics.

References

1. Zheng, X., Shen, J., Cox, C., Wakefield, J. C., Ehm, M. G., Nelson, M. R., & Weir, B. S. (2014). HIBAG - HLA genotype imputation with attribute bagging. *Pharmacogenomics Journal*, *14*(2), 192–200. https://doi.org/10.1038/tpj.2013.18

Keywords

HLA imputation, SNP, HIBAG, genomics, machine learning, immunology

OLOGRAM : Modeling the distribution of overlap length between genomic regions sets

Quentin Ferré[1,3], Guillaume Charbonnier[1], Nori Sadouni[1], Fabrice Lopez[1,5], Yasmina Kermezli[1], Salvatore Spicuglia[1,4], Cécile Capponi[3], Badih Ghattas[2] and Denis Puthier[1]

[1] Aix Marseille Univ, INSERM, UMR U1090, TAGC, Marseille, France
[2] Aix Marseille Univ, CNRS, UMR 7373, IMM, Marseille, France
[3] Aix Marseille Univ, CNRS, UMR 7020, LIS, Qarma, Marseille, France

Corresponding Author: denis.puthier@univ-amu.fr

1. Introduction

Various bioinformatics analyses can provide sets of genomic coordinates of interest. Whether two such sets possess a functional relation is a frequent question. This is often determined by interpreting the statistical significance of overlaps between the two sets [1] with binomial-based approaches such as CEAS [2], considering only peak centers and/or assessing only the number of intersections without considering their lengths. Existing methods also discard the distribution of inter-features distances while shuffling.

Here, we introduce *OLOGRAM*, which performs overlap statistics between sets of genomic regions described in BED or GTF (through features or keys). It performs Monte Carlo simulation, taking into account both the distributions of regions and inter-region lengths, to fit a Negative Binomial model of the total overlap length.

2. Results and discussion

In our model, a shuffle of a genomic region set is generated by performing independent permutations of the list of region lengths and inter-region lengths, for each chromosome. This method differs from the classical *BEDTOOLS shuffle* [3] which sets elements at random positions. Excluding regions is also possible. Computing the intersections between two shuffled sets is done using a custom algorithm of the sweep line family.

Under (H_0) (independence between the two sets), consider N the number of intersections and S the total number of overlapping base pairs ; we model both using a Negative Binomial distribution [4]. Our shuffles allow us to obtain an approximation of the distributions of these statistics under the null hypothesis, with a test to confirm the fitting.

Most of the code is written in Python 3, with performance-critical operations multi-threaded and/or written in Cython [5] This tool is available as a plugin of *pygtftk* [6], usable through its command line interface for ease of access. As such, it can be passed a GTF file treated by it (see documentation for some examples). It is available from https://github.com/dputhier/pygtftk.

References

1. Haiminen, N. et al. Determining significance of pairwise co-occurrences of events in bursty sequences. *BMC bioinformatics*, 9, 336, 2008.
2. Ji, X. et al. CEAS: cis-regulatory element annotation system. *Nucleic Acids Research*, 34, W551–W554. (2006)
3. Quinlan, A. R. and Hall, I. M. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics*, 26 (6), 841–842, 2010.
4. Omair, M. A. et al. A bivariate model based on compound negative binomial distribution. *Revista Colombiana de Estadística*, 41 (1), 87–108, 2018.
5. Behnel, S. et al. Cython: The best of both worlds. *Computing in Science Engineering*, 13 (2), 31 –39, 2011.
6. Lopez, F. et al. Explore, edit and leverage genomic annotations using python GTF toolkit. *Bioinformatics*, 2019.

Omics Data Analysis Facilities in a Biomedical Research Institute

Justine GUEGAN[1], Aurélien BELIARD[1], Mathilde BERTRAND[1], Thomas GAREAU[1], Beáta GYÖRGY[1], François-Xavier LEJEUNE[1] and Ivan MOSZER[1]

[1] Institut du Cerveau et de la Moelle épinière, Inserm U 1127, CNRS UMR 7225, Sorbonne Université, F-75013, Paris, France

Corresponding Author: justine.guegan@icm-institute.org

The iCONICS core facility is part of the Institut du Cerveau et de la Moelle épinière (ICM), which is dedicated to basic and clinical neuroscience research; it develops and makes available software solutions and methodological expertise in three domains: data management (curation, standardization, structuration, integration); high throughput genetics and omics data processing (in particular from NGS data); basic and advanced biostatistics, especially integration of multimodal data (namely clinical, omics and imaging data).

As part of the omics data analysis activity, a dedicated team within the platform assists scientific and clinical teams from the design of their study up to data processing, analysis and interpretation. This support consists of three complementary services: the building and operation of specialized pipelines to compute the raw data; the development and deployment of graphical tools to help in the interpretation of the results; a personalized assistance to biologists to go deeper in their scientific questions.

Software pipelines were built around two technologies: Snakemake [1], a workflow manager that makes pipelines scalable by enabling their parallelization; Conda [2], a package manager used to make the installation of the pipelines and their dependencies automatic. Pipelines were developed for the following types of (epi)genomics studies: gene panel and whole-exome sequencing (SNPs, CNVs, rare variants), RNA-seq (differential gene expression, fusion transcript detection, small and long non-coding RNA, single-cell), bisulfite-seq (methylation profile) [3], ATAC-seq (chromatin accessibility), and ChIP-seq (protein binding).

Shiny/R graphical applications were developed to make data available to end-users in an intuitive and interactive way. Dedicated tools are thus proposed to explore genotyping data, or transcriptomics data from RNA-seq experiments. In addition, an external software meant to filter and query genetic variants from exome data was deployed (Polyweb – Imagine, Paris Descartes), and a recent development was performed to build a database of variants identified in the frame of ICM studies, together with a graphical user interface.

Finally, *ad hoc* expertise is proposed as a follow-up of every project. Bioinformaticians in the platform dialog with biologists and clinicians to understand their scientific questions, and extract relevant information from experimental results. This activity consists in guiding scientists in the use of software and methods, and developing scripts to carry out specific processing of the data; this is tightly linked to the biostatistics component of iCONICS. Co-authored publications are a frequent outcome of such projects (e.g., [4,5]).

Acknowledgements

This work was supported by the IHU-A-ICM program ANR-10-IAIHU-06. TG is funded by the Institut Français de Bioinformatique (ANR-11-INSB-0013).

References

1. Köster J and Rahmann S. Snakemake – a scalable bioinformatics workflow engine. *Bioinformatics*, 28:2520-2522, 2012.
2. Anaconda Software Distribution. Computer software. Vers. 2-2.4.0, Nov. 2016. Web. https://anaconda.com
3. https://gitlab.icm-institute.org/iconics_public/Bistar – https://anaconda.org/icm-iconics/bistar
4. Marie C, Clavairoly A, Frah M, Hmidan H, Yan J, Zhao C, Van Steenwinckel J, Daveau R, Zalc B, Hassan B, Thomas JL, Gressens P, Ravassard P, Moszer I, Martin DM, Lu QR and Parras C. Oligodendrocyte precursor survival and differentiation requires chromatin remodeling by Chd7 and Chd8. *Proc Natl Acad Sci USA*, 115:E8246-E8255, 2018.
5. Gendron J, Colace-Sauty C, Beaume N, Cartonnet H, Guegan J, Ulveling D, Pardanaud-Glavieux C, Moszer I, Cheval H and Ravassard P. Long non-coding RNA repertoire and open chromatin regions constitute midbrain dopaminergic neuron - specific molecular signatures. *Sci Rep*, 9:1409, 2019.

Palimpsest: an R package for studying mutational and structural variants signatures along clonal evolution in cancer from single or multiple samples sequencing

Theo Z HIRSCH[1, 2], Jayendra SHINDE[1, 2], Benedict MONTEIRO[1, 2], Quentin BAYARD[1, 2], Sandrine IMBEAUD[1, 2], Feng LIU[3], Victor RENAULT[3], Jessica ZUCMAN-ROSSI[1, 2] and Eric LETOUZÉ[1, 2]

[1] Centre de Recherche des Cordeliers, Functional Genomics of Solid Tumors laboratory, Sorbonne Université, Inserm, USPC, Université Paris Descartes, Université Paris Diderot, Paris, France

[2] Labex OncoImmunology, Equipe labellisée Ligue Contre le Cancer, Centre de Recherche des Cordeliers, Paris, France

[3] Laboratory for Bioinformatics, Fondation Jean Dausset – CEPH, Paris F-75010, France

Corresponding Author: theo.hirsch@inserm.fr / eric.letouze@inserm.fr

Cancer genomes are altered by various mutational processes and, like palimpsests, bear the signatures of these different processes. Mutational signature analysis is a powerful approach to decipher the origin of somatic mutations in cancer and has been widely used since the seminal paper of Alexandrov et al in 2013 [1]. In 2018 we published the Palimpsest R package [2], which has the originality to integrate mutational signature and clonality analyses in order to reconstruct the natural history of a tumor. This combined approach allows the characterization and visualization of mutational signatures evolution during tumor development, notably the changes between early clonal and late subclonal events.

The Palimpsest package is freely available at www.github.com/FunGEST/Palimpsest. Palimpsest takes as input somatic mutations, structural variants (optional) and copy-number data obtained from whole genome or whole exome sequencing, as well as a minimal sample annotation file indicating gender, tumor purity and relation between samples (if multiple samples from a same patient were sequenced).

Beyond signatures of single base substitution, Palimpsest allows the extraction of structural variant signatures, and we are currently encoding the possibility of analyzing signatures of doublet base substitutions as well as small insertions and deletions (indels). For all those signatures, the user has the choice between the *de novo* extraction of novel signatures and the quantification of previously described signatures using non-negative matrix factorization. Palimpsest also estimates the probability of each mutation being due to each process to predict the mechanisms at the origin of driver events.

We are currently integrating new features in the Palimpsest package in order to decipher the clonal architecture of a tumor using sequencing data of multiple samples from the same patient. This analysis relies on a Bayesian Dirichlet process for defining clusters of mutations corresponding to different clones, an approach already used for studying synchronous [3] or metachronous [4] multisampling. Those clones can later be used for constructing a phylogenic tree of the tumor, revealing the evolution and diversification of mutational processes during the natural history of the tumor.

References

1. Ludmil B. Alexandrov et al. Signatures of mutational processes in human cancer. *Nature*, 500, 415–421, 2013.
2. Jayendra Shinde et al. Palimpsest: an R package for studying mutational and structural variant signatures along clonal evolution in cancer. *Bioinformatics*, 34(19), 3380-3381, 2018.
3. Lucy R. Yates et al. Subclonal diversification of primary breast cancer revealed by multiregion sequencing. *Nature Medicine*, 21(7), 751–9, 2015.
4. Gunes Gundem et al. The evolutionary history of lethal metastatic prostate cancer. *Nature*, 520(7547), 353–7, 2015.

PanGBank: depicting microbial species diversity via PPanGGOLiN

Guillaume Gautreau[1], Adelme Bazin[1], Rémi Planel[1], Mathieu Gachet[1], Mathieu Dubois[1], Laura Burlot[1], Amandine Perrin[2,3] Marie Touchon[2,3], Eduardo Rocha[2,3], Christophe Ambroise[4,5], Catherine Matias[6], Claudine Medigue[1] and David Vallenet[1]

[1] LABGeM, Génomique Métabolique, CEA, Genoscope, Institut François Jacob, CNRS, Université d'Évry, Université Paris-Saclay, Evry, France
[2] Microbial Evolutionary Genomics Unit, Institut Pasteur, 75724 Paris, France
[3] CNRS-UMR-3525, 75015 Paris, France
[4] Laboratoire de Mathématiques et Modélisation d'Evry (LaMME), 91000 Evry, France
[5] UMR-CNRS-8071, 91000 Evry, France
[6] Laboratoire de Probabilités et Modèles Aléatoires (LPMA), 75252 Paris, France

Corresponding Author: ggautrea@genoscope.cns.fr

By collecting and comparing genomic sequences, many studies are focused on the overall gene content of a species (*i.e.* the pangenome [1]) to understand its evolution in terms of core and accessory parts. The core genome is defined as the set of genes shared by all the organisms of a taxonomic unit (generally a species). Accessory part (variable regions) is crucial to understand the adaptive potential of bacteria and contains genomic regions that are exchanged between strains by horizontal gene transfer (HGT). However, this dichotomy is not robust against poorly sampled data because it is highly reliant on the presence or absence of a single organism and also does not faithfully report the diverse ranges of gene frequencies in a pangenome. Moreover, this approach considers genomes as isolated gene sets and neglects their chromosomal organization despite its major importance to study HGT. Here, we introduce a compact modelization of multiple genomes, giving it a representation using a graph model built up from genes clustered into gene families coupled with a statistical partitioning method.

The PPanGGOLiN method merges the chromosomal links between neighboring genes to build a graph of the neighborhood between gene families weighted by the number of genomes covering each edge. In addition to the graph, the pangenome is modeled as a binary presence/absence matrix where the rows correspond to gene families and columns to the genomes (1 in case of presence of at least one gene belonging to this gene family, 0 in case of absence). The pangenome is then partitioned by evaluating, through an Expectation-Maximisation algorithm, the best parameters of a Bernoulli Mixture Model (BMM). This approach partitions pangenomes into three types of genomes: (1) *persistent genome*, equivalent to a relaxed core genome (genes conserved in all but a few genomes); (2) *shell genome*, genes having intermediate frequencies corresponding to moderately conserved genes potentially associated with environmental adaptation capabilities; (3) *cloud genome*, genes found at very low frequency. Finally, the partitions are overlaid on the neighborhood graph to obtain a Partitioned Pangenome Graph (PPG).

This method was applied on all the genomes available in the GenBank database (encompassing ~600 species and ~200 000 genomes) to obtain a database of PPGs. This in-development resource, called PanGBank, provides a wide view of the different range of gene frequencies and chromosomal topologies along the microbial world thanks to an API and a web visualization tool dedicated for browsing PPGs. In the context of massive comparative genomics, drawing genomes on rails like a subway map may help biologists to compare their genomes of interest to the overall pangenomic diversity.

References

[1] Tettelin, H., Masignani, V., Cieslewicz, M. J., Donati, C., Medini, D., Ward, N. L., ... & DeBoy, R. T. Genome analysis of multiple pathogenic isolates of Streptococcus agalactiae: implications for the microbial "pan-genome". *Proceedings of the National Academy of Sciences*, *102*(39)13950-13955. 2005.

[2] Ambroise C., Dang M. and Govaert G. Clustering of spatial data by the EM algorithm. *geoENV-I-Geostatistics for environmental applications*, pages 493-504, 1997.

Pathway analysis from time course gene expression experiments to unveil the dynamic of cellular responses

Michaël PIERRELÉE[1], Laurent TICHIT[2] and Bianca HABERMANN[1]

[1] Aix Marseille Univ, CNRS, IBDM, Campus de Luminy, 13009 Marseille, France
[2] Aix Marseille Univ, CNRS, I2M, Campus de Luminy, 13009 Marseille, France

Corresponding Author: michael.pierrelee@univ-amu.fr

1 Introduction

According to a common definition, a « biological pathway is a series of actions among molecules in a cell that leads to a certain product or a change in the cell » [1]. In other words, a stimulus leads to actions within the cell, then to cellular responses. A response to a stimulus comes from interactions between molecules within the cell. The former is modeled as edges and the latter as nodes, shaping a network. Active nodes have a property changing over time and they affect the neighborhood nodes through the edges - the interactions. This dynamic has to be taken into account to fully understand cellular mechanisms.

Active biological pathway can be found using RNA-sequencing. In a time-course experiment, it takes a series of snapshots of gene expression, allowing to detect genes only differentially expressed at a given time and which would be ignored otherwise.

Several algorithms exist to exploit these multidimensional data. A standard approach is to calculate gene-wise correlations and draw a network from them [2]; but this model fails to represent the network dynamic. To overcome that, a method was designed to categorize genes according to the time of their highest fold change and then find a path from the early to late ones [3]. However, this algorithm excludes by definition genes which are involved at different time points in pathways.

2 Temporal network project

In this project, we will use multilayer networks to generate a temporal network representing the evolution of the cellular response over time. The interaction network will be weighted differently at each time, depending on the differential expression over time. It will give a set of networks, each one corresponding to a snapshot of the cell changes at a given time-point. Thus, there is a causality from one time to the next. Multilayer network approaches allow to represent this causality by a directed “inter-layer edge” from a node at a given time to itself at the next time [4]. The interaction network will come from high quality interaction databases. For development purpose, the model organism will be the yeast and the network will be limited to protein-protein interactions, available on HitPredict [5], involved in yeast cell cycle. We will then adapt network analysis to extract optimal active subnetworks from the temporal network.

3 Conclusion

The understanding of cellular mechanisms involves taking into account the whole dynamics of cellular changes over time. Temporal networks are an original way to model the different states of nodes over time, compared to standard approaches which consider that a node is intrinsically the same at each time.

References

[1] National Human Genome Research Institute (NHGRI). (2019). Biological Pathways Fact Sheet. [online] Available at: https://www.genome.gov/27530687/ [Accessed 13 Mar. 2019].
[2] van Dam, S., et al. (2017). Gene co-expression analysis for functional classification and gene–disease predictions. Brief. Bioinform. 19, bbw139.
[3] Patil, A., and Nakai, K. (2014). TimeXNet: identifying active gene sub-networks using time-course gene expression profiles. BMC Syst. Biol. 8, S2.
[4] Kivelä, M., et al. (2014). Multilayer networks. J. Complex Networks 2, 203–271.
[5] López, Y., et al. (2015). HitPredict version 4: Comprehensive reliability scoring of physical protein-protein interactions from more than 100 species. Database 2015, bav117.

Performance evaluation of bioinformatics tools for predicting allergenic proteins in food

Wafa Mokhtari[1,2], Souad Khemili-Talbi[1], Jean Marc Kwasigroch[2] and Dimitri Gilis[2]

[1] Laboratory VALCORE, M'Hamed Bougara University of Boumerdès, 35000, Boumerdès, Algeria.
[2] Laboratory 3BIO-BioInfo, Université libre de Bruxelles, 1050, Brussels, Belgium.

Corresponding Author: w.mokhtari@univ-boumerdes.dz , Wafa.Mokhtari@ulb.ac.be

Food allergy is a health problem due to proteins that are not tolerated by the immune system. Predicting whether a protein used in food could be an allergen is a crucial issue. The FAO has developed guidelines for evaluating the potential allergenicity of proteins used in diet: a protein is considered as a potential allergen if it shares with a known allergen more that 35% sequence identity on a window of 80 amino acids or a perfect identity with 6 contiguous amino acids [1]. Several other bioinformatics tools have been developed to predict the allergenicity of a protein on the basis of different sequence/structure features, which are combined with statistical or machine learning methods.

We evaluated the performance of different tools that predict allergenic proteins in food and that are available on a webserver or as a standalone program; one of them is based on the application of the FAO rules. We created a database of food allergens and non-allergens. We used this database to evaluate the performance of the considered prediction programs. When all the possible biases between our evaluation database and the learning sets of the tested tools are removed, the application of the FAO rules leads to the best scores. But this method suffers from a very large number of false positives. As a perspective, we aim at identifying features that could be combined with the FAO rules to decrease the number of false positives.

Keywords: Bioinformatics, Food allergen, Prediction.

References

1. M WEJ Fiers, G A Kleter, H Nijland, A A Peijnenburg, J P Nap and R CHJ van Ham. Allermatch™, a webtool for the prediction of potential allergenicity according to current FAO/WHO Codex alimentarius guidelines. BMC Bioinformatics, 133-138, 2004.

Pioneer data-driven methods generating synthetic data: the HLA "avatars" are shifting paradigms in data sharing.

Estelle GEFFARD[1,2], Thomas GORONFLOT[3], Sophie LIMOU[1,2,4], Nicolas VINCE[1,2], Matthieu WARGNY[3], Pierre-Antoine GOURRAUD[1,2,3]

[1] ATIP-Avenir, Centre de Recherche en Transplantation et Immunologie UMR 1064, INSERM, Université de Nantes, Nantes, France

[2] Institut de Transplantation Urologie Néphrologie (ITUN), CHU Nantes, Nantes, France

[3] CHU de Nantes, INSERM, CIC 1413, Pôle Hospitalo-Universitaire 11 : Santé Publique, Clinique des données, Nantes, France

[4] Ecole Centrale de Nantes, Nantes, France

Corresponding Author: pierre-antoine.gourraud@univ-nantes.fr

The enforcement of the GDPR regulation has shed a new light on data protection and privacy issues in both care and research. Risk of re-identification is more than ever a central concern for all European regulations often translating into constraints for data-sharing in science. GDPR may thus impact research reproducibility in science, data-sharing efforts and ultimately data-driven care for patients. We propose a simple solution to share individual HLA genotypes data without compromising on privacy: generate HLA "avatars" from real HLA genotypes data. Based on combination of founder haplotypes estimated by an EM-algorithm from HLA genotypes, under Hardy Weinberg Equilibrium (HWE) proportions, we use an in-silico genetic resampling of HLA haplotypes to generate HLA genotypes of unidentifiable virtual individuals: "avatars". These HLA genotypes must preserve the individual structure of the original dataset and keeps unchanged global parameters such as allele frequencies, genotype frequencies, sum of top 10, 25, 50 haplotype frequencies (respectively ~20%, 25-30%, ~35% in a population of European ancestry). However, because the "avatarization" process may mimic evolutive bottleneck, the total number of haplotypes is reduced in a statistically significant log-linear dependent way to the sample size ($p<10^{-4}$). Haplotypes and alleles occurring less than 5 times in the original dataset are prone to over -and under- sampling, as anticipated by the Gaussian normality approximation ($n*p*(1-p) > 5$). Avatarization can be improved by informed iterative resampling that corrects the natural sampling truncation of HLA haplotype under HWE. Beyond genetics, this "digitally-assisted in silico procreation" is a promising data-driven way to facilitate data sharing. The resampling method can also accommodate clinical and demographic annotations by stratification calling for a generalized framework to create avatars in data sharing and data governance in the post-GDPR era context.

Keywords

HLA, anonymization, GDPR, statistical avatars

SChnurR : Pipeline d'analyse et de visualisation scRNA-seq

Jonathan CRUARD[1], Mathias BAGUENEAU[1], Jean-Baptiste ALBERGE[1] and Stéphane MINVIELLE[1,2]

[1] CRCINA, INSERM, CNRS, Université de Nantes, Université d'Angers, 8 quai Moncousu, 44007, Nantes, France

[2] CHU de Nantes, 5 allée de l'île gloriette, 44093, Nantes, France

Corresponding Author: stephane.minvielle@chu-nantes.fr

1 Introduction

Le séquençage d'ARN en cellule unique (scRNA-seq) apporte des informations nouvelles sur la diversité des sous-populations de cellules de multiples tissus et contribue à des découvertes scientifiques dans le développement, l'immunité et la recherche contre le cancer. L'analyse de l'hétérogénéité d'une expérience de scRNA-seq exploite une matrice de comptage d'expression N (cellules) x P (gènes) dont on cherche à extraire des motifs (similarité et variabilité entre cellules).

Des artefacts techniques identifiés (encapsulation de doublets de cellules, contamination d'ARNm, etc.) impactent l'étude de l'expression et la comparaison de sous-populations cellulaires. Ces artefacts masquent la réalité biologique de l'hétérogénéité des cellules uniques.

2 Pipeline

Nous proposons d'intégrer dans le pipeline d'analyse SChnurR des outils de réduction de biais techniques et de clustering. SoupX [1] : cet algorithme tend à corriger la part d'expression provenant de la contamination du milieu pour chaque cellule. Celui-ci estime le profil de la soupe à partir de la matrice de comptage brute (capsules « vides » incluses), puis mesure la part de contamination pour chaque cellule. Enfin à partir de ces résultats le profil d'expression de chaque cellule est corrigé. SCDS [2] ce package affecte un score à chaque cellule à partir de 2 méthodes : la première se base sur les co-expressions observées de chaque paire de gènes. Le second est basée sur la construction de doublets artificiels à partir des données, l'algorithme apprend alors à les différencier. À partir de cet apprentissage l'algorithme évalue la probabilité de chaque vraie cellule d'être en réalité un doublet. Le score obtenu résume alors les résultats de chaque approche. Sctransform [3] : la transformation appliqué par ce package prend en compte les biais techniques et permet d'éviter l'overfitting induit par les modèles basés sur la distribution binomiale négative. Le pipeline développé est basé sur le package Seurat [4], celui-ci fournis des fonctionnalités de visualisation et sert d'interface avec les différents packages.

3 Conclusion

Les sous-populations artefactuelles composées de doublets sont supprimées par le pipeline. La normalisation des ARN contaminants corrige les tests d'expression différentielle entre plusieurs expériences. Dans un second temps une application Rshiny a été développée afin de permettre la visualisation des jeux de données corrigés et normalisés.

References

[1] Matthew Daniel Young (2018). SoupX: Single cell mRNA Soup eXterminator. R packageversion 0.3.0.

[2] AS Bais, D Kostkam (2019). scds: Computational Annotation of Doublets in Single Cell RNA Sequending Data. bioRxiv. URL https://doi.org/10.1101/564021

[3] Christoph Hafemeister (2018). sctransform: Variance Stabilizing Transformations for Single Cell UMI Data. R package version 0.1.0. https://CRAN.R-project.org/package=sctransform

[4] Butler et al. Integrating single-cell transcriptomic data across different conditions, technologies, and species. Nature Biotechnology (2018).

Polygenic Risk Scores for Autism spectrum disorder and Alzheimer's disease enable the identification of new white matter tract biomarkers

Thomas RIQUELME[1], Antoine GRIGIS[1], Cathy PHILIPPE[1] and Vincent FROUIN[1]

[1] Neurospin, Institut Joliot, CEA, 91191, Gif-sur-Yvette, France.

Corresponding Author: vincent.frouin@cea.fr

1 Introduction

A polygenic risk score is a cumulative genetic risk computed with one subject's genome variants. Polygenic risk scores have gained interest as they can be correlated to a newly available phenotype in an independant cohort, and offer a mean to detect shared etiology between traits [1]. In this study, we computed polygenic risk scores for Autism Spectrum Disorder and Alzheimer's disease for healthy subjects in two large imaging-genetic cohorts, UK Biobank and Human Connectome Project. Then, we assessed the correlations between these scores and brain imaging derived phenotypes obtained in 48 white matter tracts. We aimed to identify potential white matter tract biomarkers for Autism and Alzheimer's disease.

2 Methods

In this work, we used the two largest imaging-genetics cohorts available as open data to date: Human Connectome Project (820 subjects; dbGaP appl. #17771) and UK Biobank (14,538 subjects; appl. #25251). The summary statistics for Autism and Alzheimer's disease were retrieved from two large Genome-Wide Association Studies [2,3]. PRSice tool [1] was then used to compute polygenic risk scores for Autism and Alzheimer's disease for each subject in both cohorts using the summary statistics of each diseases. As new phenotypes, we considered measures on the main human white matter tracts: namely, we computed the average Fractional Anisotropy in 48 white matter tract masks obtained from the JHU atlas [4]. Afterward, PRSice was used to compute associations between polygenic risk scores and white matter tract measurements. False discovery rate (FDR) was used to correct for multiple testing.

3 Results and discussion

For both conditions, we found about ten significant associations between white matter tracts and polygenic risk scores of the diseases (FDR < 0.01). The most significant tracts that replicated in both cohorts were respectively the superior corona radiata for Autism and the middle cerebellar peduncle for Alzheimer's disease. Some already known associations were recovered [5,6]. Especially, Pryweller et al. suggested that the superior corona radiata contains motor and sensory fibers projecting to cortex whose alteration could cause hypo- or hyper-responsiveness in Autism; and Miyasaka et al. explained that the middle cerebellar peduncle contains afferent fibers from the pons to the cerebellum which support the hypothesis of a role of the cerebellum in Alzheimer's disease. These white matter tracts could be important manifestation of the genetic predispositions for Autism and Alzheimer's disease and thus could represent potential biomarkers for these conditions. This work illustrates how polygenic risk score analysis may help in detecting the brain structures in which the genetic predisposition for a syndrome manifests itself in the general population.

References

[1] Jack Euesden, Cathryn M. Lewis, and Paul F. O'reilly. PRSice: polygenic risk score software. *Bioinformatics*, 31.9: 1466-1468, 2014.

[2] Jakob Grove et al. Identification of common genetic risk variants for autism spectrum disorder. *Nature genetics*, 1, 2019.

[3] Jean-Charles Lambert et al. Meta-analysis of 74,046 individuals identifies 11 new susceptibility loci for Alzheimer's disease. *Nature Genetics*, 45: 1452–1458, 2013.

[4] Susumu Mori et al. MRI atlas of human white matter. Elsevier, 2005.

[5] Jennifer R. Pryweller et al. White matter correlates of sensory processing in autism spectrum disorders. NeuroImage: Clinical, 6:379-387, 2014.

[6] Toshiteru Miyasaka et al., Cerebellar White Matter Involvement in Alzheimer's Disease: Diffusion Tensor Study. Radiological Society of North America 2014 Scientific Assembly and Annual Meeting, - ,Chicago IL

Population demographic estimation using simulated data

Elisabeth QUILLERY[1], Charlotts BERTHELLIER[1], Christian DINA[1] and Isabel ALVES[1]

[1] l'institut du thorax, INSERM, CNRS, Univ Nantes, CHU Nantes, Nantes, France

Corresponding Author: Christian.dina@univ-nantes.fr

Broadly speaking, population genetics can be defined as the study of the genetical basis of naturally occurring variation, with the aim of describing and understanding the evolutionary forces that create variation within species and which lead to differences between species. Studies in this branch of biology examine such phenomena as adaptation, speciation, and population structure.

For the purposes of our work, we'll be focusing on population structure. From the viewpoint of population geneticists, population structure takes into account the organization of genetic variation within and between populations, with special emphasis on their spatial arrangement. It is the presence of a systematic difference in allele frequencies between sub-populations in a population, possibly due to different ancestry.

In this work, our aim is to test the reliability and robustness of two methods which estimate population size evolution in the past from present genetic data. These two methods are respectively based Identity by Descent [2] or on the Site Frequency Spectrum [3] information. The objective is testing hypothesis of evolutionary models by contrasting and inferring demographic models and see the impact of population migration or geographical barriers on population size evolution estimates. This pipeline will be used later by the teams in the laboratory.

We begin by running a whole genome simulation using the ARGON [4] software (ARGON.0.1.jar), which is a simulator for the discrete time Wright-Fisher model (DTWF) process. We automatically generate the population model that ARGON will then use to run the simulation at the genome level.

In this first example, the evolution model we use represents the case of a population split scenario, and starts with 2 populations with 10000 individual each.

We are applying traditional quality control analyses. When the data has been effectively pruned we can then create an IBS (identity by state) matrix, which measures the proportion of sites at which 2 chromosomes are the same. And this is done for all chromosomes for each individual and compares each individual 2 by 2 to establish a similarity matrix. Lastly we proceed with the –mds-plot option of plink in order to run multidimensional scaling on our data which can then be plotted in R (R/3.3.3) and we can see how the individual in our data tend to cluster together. Finally, we could apply the two methods based on IBD information [2] and SFS [3] to evaluate the effect of this admixture on estimated population size changes.

Our initial results showed that we have the power to detect a true bottleneck, brutal decrease of population size, when it happens and that this event can be discovered whatever the chunk size of the identity by descent. We are now extending the model in order to evaluate the proportion of falsely inferred bottleneck both in presence of no event at all or in presence of simple admixture.

The present pipeline allows us to test the reliability of existing methods for populations size estimation and may be developed in order to evaluate more complex models as past admixture.

References

1. Mathieson et al. « Eight Thousand Years of Natural Selection in Europe ». *BioRxiv*, 10 octobre 2015. https://doi.org/10.1101/016477.
2. Browning, et al. Accurate non-parametric estimation of recent effective population size from segments of identity by descent. The American Journal of Human Genetics 97,
3. Excoffier, et al. Robust demographic inference from genomic and snp data. PLoS genetics 9, 10 (2013), e1003905.
4. Palamara, P. F. Argon: fast, whole-genome simulation of the discrete time wright-fisher process. Bioinformatics 32, 19 (2016), 3032–3034.

Positive Multistate Protein Design

Jelena Vucinic[1,2], David Simoncini[1,3], Manon Ruffini[1,2], Sophie Barbe[1] and Thomas Schiex[2]
[1] LISBP, Université de Toulouse, CNRS, INRA, INSA, Toulouse, France
[2] MIAT, Université de Toulouse, INRA, Auzeville-Tolosane, France
[3] IRIT UMR 5505-CNRS, Université de Toulouse, 31042 Cedex 9, France.

Corresponding author: thomas.schiex@inra.fr

Reference paper: **Vucinic *et al.* (2019) Positive Multistate Protein Design. Submitted.**

Abstract *Structure-based Computational Protein design (CPD) plays a critical role in advancing the field of protein engineering. Using an all-atom energy function, CPD tries to identify amino acid sequences that fold into a target structure and ultimately perform a desired function. The usual approach considers a single rigid backbone as a target which ignores backbone flexibility. Multistate design (MSD) allows instead to consider several backbone states simultaneously. The paper presents two reductions of positive multistate protein design to Cost Function Networks that outperform state-of-the-art guaranteed computational design approaches by orders of magnitudes and can solve MSD problems with sizes previously unreachable with guaranteed algorithms.*

Keywords Computational protein design, graphical models, discrete optimization

Computational Protein Design (CPD) seeks to identify sequences of amino acids that adopt a desired tertiary structure and ultimately performs a desired function. It requires an energy function that reflects protein stability and a search method to identify a sequence with a conformation of optimal stability. Because of the intractable combination of the many degrees of freedom of a protein and the non-convex form of even the crudest energy functions, this problem has been simplified by several assumptions: the energy is supposed to be described as a pairwise decomposable function, the protein backbone degrees of freedom are fixed to an idealized target backbone and the side-chain of each amino acid is assumed to adopt one of a finite set of possible conformations or rotamers.

Despite these simplifications, the size of the search space remains exponentially large and the problem of searching for a sequence with a minimum energy conformation is decision NP-complete [1]. Therefore, most CPD approaches rely on stochastic optimization algorithms such as Monte Carlo Simulated Annealing or Genetic algorithms, which provide only asymptotic convergence guarantees. Recent progress in guaranteed discrete optimization showed that stochastic methods may durably fail to find or get close to the optimum when the problem becomes hard. Despite years of CPU-time, a tuned Simulated Annealing algorithm was unable to find the global energy optima that was identified and proved as optimal by Cost Function Networks (CFN) algorithms [2]. The recent design of the hyper-stable self-assembling β-propeller "Ika" by CFN technology [3] shows that guaranteed methods can also be useful in practice, combining efficiency with the assurance that optimization didn't fail.

In the paper, we combine the guarantees and efficiency of CFN algorithms with the idea of defining the target structure as an ensemble of backbones conformations instead of a single idealized structure. Compared to the usual SSD approach, multistate design (MSD) has shown to provide enhanced design capacities to stabilize an ensemble of backbones or proteins with specific binding properties [4]. In these cases, MSD seeks to identify a sequence that optimizes a function of its optimal energies on the different considered states. This function, or "fitness", is itself non trivial to compute, as it requires the computation of optimal conformations of the sequence on several backbone states. Many SSD optimization algorithms have been extended to MSD, with more or less general fitness functions, including Monte Carlo with simulated annealing [5], genetic algorithms [6], cluster expansion [4], and dead-end-elimination in combination with A* [8].

The nature of the fitness function intimately depends on the design problem. When the aim is to design a sequence that fits several conformational states, the fitness will typically be the average of the energies on all states, a typical example of *positive* multistate design. For specificity however,

undesirable states are present and the fitness function would be defined as the difference in energy between desirable (positive) and undesirable (negative) states.

In our paper, we showed that the fitness function used has a profound influence on the computational nature of the problem. Specifically, the introduction of negative states makes the problems qualitatively more complex and precisely $\mathsf{NP}^{\mathsf{NP}}$-complete [9]. This result has several implications. Negative MSD being harder than SSD, optimization methods may become unable to reach good quality solutions sooner than in the SSD case. It also shows that positive MSD interesting as it is "just" NP-complete while capturing some backbone flexibility. Hence, we leverage the polynomial equivalence of NP-complete problems by introducing efficient reductions of two variants of positive multistate design to Cost Function Networks. The first variant uses a (weighted) average energy fitness and the second one a minimum energy fitness. Beyond saving programming efforts, this approach directly benefits from the advanced CFN processing machinery [10].

On various positive MSD problems, we showed that it is possible to identify an optimal MSD sequence with associated optimal conformations in reasonable time, on computationally extremely challenging design problems of a size far beyond what has been solved with existing state-of-the-art guaranteed multistate design methods [8], including recent CFN based methods with dedicated algorithms [11]. Our software is also natively able to exhaustively enumerate suboptimal sequences close to the MSD optimum, which is convenient for sequence library design. Contrarily to what has been previously described [12], we observe that the use of an ensemble of NMR structures as a positive ensemble of backbones provides strong improvements in term of native sequence and sequence similarity recovery when an average energy criteria is used. We also show that this improvement is reduced but still present when a backrub generated ensemble derived from a single X-ray structure is used. These results show that Positive Multistate Design is essentially as hard to solve as Single State Design, both in theory and in practice. Given the significant improvement that the multistate approach brings, positive MSD should be considered as a default approach when specificity is not the main target.

References

[1] Niles A Pierce and Erik Winfree. Protein design is np-hard. *Protein engineering*, 15(10):779–782, 2002.

[2] David Simoncini, David Allouche, Simon de Givry, Céline Delmas, Sophie Barbe, and Thomas Schiex. Guaranteed discrete energy optimization on large protein design problems. *Journal of chemical theory and computation*, 11(12):5980–5989, 2015.

[3] Hiroki Noguchi, Christine Addy, David Simoncini, Staf Wouters, Bram Mylemans, Luc Van Meervelt, Thomas Schiex, Kam YJ Zhang, JRH Tame, and ARD Voet. Computational design of symmetrical eight-bladed β-propeller proteins. *IUCrJ*, 6(1), 2019.

[4] Christopher Negron and Amy E Keating. Multistate protein design using clever and classy. In *Methods in enzymology*, volume 523, pages 171–190. Elsevier, 2013.

[5] Xavier I Ambroggio and Brian Kuhlman. Computational design of a single amino acid sequence that can switch between two distinct protein folds. *Journal of the American Chemical Society*, 128(4):1154–1161, 2006.

[6] Navin Pokala and Tracy M Handel. Energy functions for protein design: adjustment with protein–protein complex affinities, models for the unfolded state, and negative design of solubility and specificity. *Journal of molecular biology*, 347(1):203–227, 2005.

[7] Mark A Hallen and Bruce R Donald. Comets (constrained optimization of multistate energies by tree search): A provable and efficient protein design algorithm to optimize binding affinity and specificity with respect to sequence. *Journal of Computational Biology*, 23(5):311–321, 2016.

[8] Larry J Stockmeyer. The polynomial-time hierarchy. *Theoretical Computer Science*, 3(1):1–22, 1976.

[9] Barry Hurley, Barry O'Sullivan, David Allouche, George Katsirelos, Thomas Schiex, Matthias Zytnicki, and Simon De Givry. Multi-language evaluation of exact solvers in graphical model discrete optimization. *Constraints*, 21(3):413–434, 2016.

[10] Mostafa Karimi and Yang Shen. iCFN: an efficient exact algorithm for multistate protein design. *Bioinformatics*, 34(17):i811–i820, 2018.

[11] Benjamin D Allen, Alex Nisthal, and Stephen L Mayo. Experimental library screening demonstrates the successful application of computational protein design to large structural ensembles. *Proceedings of the National Academy of Sciences*, 2010.

Predicting isoform transcripts: lessons from human, mouse and dog

Nicolas Guillaudeux[1], Catherine Belleannée[1], Samuel Blanquart[1] and Jean-Stéphane Varré[2]
[1] Univ Rennes, Inria, CNRS, IRISA, Rennes F-35000, France
[2] Univ. Lille, CNRS, Centrale Lille, UMR 9189 - CRIStAL - Centre de Recherche en Informatique Signal et Automatique de Lille, F-59000 Lille, France

Corresponding author: nicolas.guillaudeux@inria.fr

1 Transcript prediction using comparative genomics

Several mechanisms, including alternative transcription and alternative splicing, enable an eukaryotic gene to express a large diversity of products [1]. The latter process allows various isoform transcripts to be built, each one made of a specific combination of genomic segments: the exons. We currently do not know how to determine the whole catalog of isoform transcripts that can be expressed from a gene. Particularly, RNA-seq data allows us to identify only a subset of the expressed transcripts, for technical reasons and due to the low expression levels of some transcripts [2,3].

To fill in the knowledge about transcript isoforms expressed from a gene, we have proposed a comparative genomics method allowing to identify orthologous exons shared by a pair of genes [4]. This method uses functional sites (start, stop codons and splice sites) known in a given source gene to transpose them, through sequence homology search, into the target orthologous gene. From orthologous exons thus identified, it is possible to estimate whether a transcript of the source gene has a splicing ortholog, i.e. whether its exon combination can also be expressed by the target gene. The method has been validated on orthologous genes shared by human and mouse [4]. In this work, we adapt the approach for a multi-species comparison and we predict transcriptomes in human, mouse, and in a non-model organism, the dog.

2 Multi-species comparison: human, mouse, dog

We analyzed 2,167 gene orthologs in human, mouse and dog, as well as a total of 18,109 known isoform transcripts (derived from CCDS for human and mouse, and Ensembl for dog). From these data, 6,861 new transcripts were predicted, thus adding to the known transcriptomes 15.5%, 24.5% and 50% putative new transcripts in human, mouse and dog, respectively. The majority of the predictions concern the dog, a non-model species, obviously less documented. Some of the predictions made in dog and mouse were validated using experimental data (RNA-seq Illumina and Oxford Nanopore) and annotations missing from CCDS and Ensembl databases.

In order to perform multi-gene comparisons from pairwise gene comparisons, we defined a graph: functional sites, known or predicted, correspond to the graph vertices, while an edge indicate an estimated orthology between two sites. From the 2,167 orthologous genes, we identified a subset of 135 genes with a structure conserved in the three species. Each of these genes shares the same functional sites as well as the same transcript potential: any transcript observed in a species is syntactically expressible in the other two. These observations can lead to phylogenetic interpretations. A first one suggests that in the ancestor of *Boreoeutheria*, each of these 135 genes already possessed the intron/exon structure conserved in human, mouse and dog; a second one indicates that the majority of the 2,167 genes examined have diverged in their intron/exon structure since their common ancestor.

References

[1] F. E Baralle and J. Giudice. Alternative splicing as a regulator of development and tissue identity. *Nature reviews. Molecular cell biology*, 18(7):437–451, 2017.

[2] T. Steijger, J.F. Abril, P.G. Engström, F. Kokocinski, RGASP Consortium, T.J. Hubbard, R. Guigó, J. Harrow, and P. Bertone. Assessment of transcript reconstruction methods for RNA-seq. *Nature methods*, 10(12):1177–1184, 2013.

[3] K. Križanovic, A. Echchiki, J. Roux, and M. Sikic. Evaluation of tools for long read RNA-seq splice-aware alignment. *Bioinformatics (Oxford, England)*, 34(5):748–754, 2018.

[4] S. Blanquart, J.S. Varré, P. Guertin, A. Perrin, A. Bergeron, and K.M. Swenson. Assisted transcriptome reconstruction and splicing orthology. *BMC Genomics*, 17(786), 2016.

Prediction of candidate disease genes through deep learning on multiplex biological networks

Stefani DRITSA[1,2,§], Thibaud MARTINEZ [1,3], Weiyi ZHANG[1,3,6], Chloé-Agathe AZENCOTT[3,4,5], Antonio RAUSELL[1,2,*]

[1] Clinical Bioinformatics Laboratory, Imagine Institute, Imagine Institute, Paris Descartes University, Sorbonne Paris Cité, 75015 Paris, France
[2] INSERM UMR 1163, Institut Imagine, 75015 Paris, France
[3] MINES ParisTech, PSL Research University, CBIO-Centre for Computational Biology, 75006, Paris, France
[4] Institut Curie, PSL Research University, 75005, Paris, France,
[5] INSERM, U900, 75005, Paris, France
[6] Shanghai Jiao Tong University

(§) Presenting author
(*) Corresponding Author: antonio.rausell@inserm.fr

The causal genes or variants of around 50% of all known Mendelian diseases described to date have still not been identified. Network propagation approaches using biological networks, including the human interactome, regulome, phenome and diseasome, have been successfully contributed to the discovery rate on new disease genes [1]. In this study we propose to leverage recent deep learning advances in semi-supervised node labeling to address multiplex network integration in disease gene discovery.

To that aim we implemented a model that, taking a given number of networks and a node class as input, uses both unsupervised learned embeddings [2] and supervised graph-structure learning. Thus, for a collection of (un)directed and (un)weighted graphs $G=(V,E,R)$, with nodes $v_i \in V$, edges $(v_i,r,v_j) \in E$ of relation type $r \in R$ (number of nodes $|V|=N$, number of relations $|R|=R$), we perform node representation learning through *node2vec* and run normalized Relational graph convolutional networks (*R-GCNs,* [3]) on all networks. Algorithmic novelties to incorporate node attributes, attention mechanisms, and unsupervised clustering into the learning process have been developed.

The model uses as input a collection of more than 100 biological networks, including protein-protein interactions, tissue-specific gene regulatory and co-expression networks, signaling networks, and functional similarity networks based on different ontologies. It was hence tested by its ability to prioritize Mendelian diseases genes of different categories and benchmarked against reference state-of-the-art methods. Detailed examples are presented and results interpreted in the context of the associated local network topologies.

References

1. Lenore C, Ideker T, Raphael BJ, Sharan R. Network propagation: a universal amplifier of genetic associations. *Nature Reviews Genetics*, 18, 551–562, 2017.
2. Grover, A.,and Leskovec J. node2vec: Scalable feature learning for network. *ACM*, 2016.
3. Schlichtkrull, M., Kipf, T. N., Bloem, P., Van Den Berg, R., Titov, I., and Welling, M. Modeling relational data with graph convolutional networks. In *European Semantic Web Conference*, pp. 593-607. Springer, Cham., 2018.

PREDIdicting bacterial PATHogenicity on plant: PREDIPATH

Felipe LIRA[1], Gilles HUNAULT[2], Martial BRIAND[1], Perrine PORTIER[1], Claudine LANDES[1] and Marion FISCHER-LE SAUX[1]

[1]IRHS, INRA, Université d'Angers, Agrocampus-Ouest, SFR 4207 QuaSaV, 49071, Beaucouzé, France.

[2]Hémodynamique, Interaction Fibrose et Invasivité Tumorales Hépatiques Laboratory, Unité Propre de Recherche de l'Enseignement Supérieur 3859,Structure Fédérative de Recherche 4208,Bretagne Loire University, Angers, France.

Corresponding Author: felipelira3@gmail.com

Currently, the prediction of bacterial pathogenicity relies on traditional microbiological methods which are time consuming and require specialists. Thus, comparative genomics emerges as a promising method which allows to check *in silico* the presence of genomics elements to distinguish pathogenic (P) from non-pathogenic (NP) organisms. In order to predict the potential pathogenicity of plant associated bacteria, we designed the PREDIPATH workflow to detect genomic markers associated to bacterial phenotypes.

Three strategies were used : I – An *a priori* approach consisted in detecting potentially over-represented genes in pathogens. For that purpose, the PREDIPATH database, composed of genes collected from public repositories and involved in virulence [1], and antimicrobial, biocides and heavy metal resistance [2,3] was created. Biosynthetic gene clusters encoding the production of secondary metabolites were searched using antiSMASH 4 database [4]; II – A without *a priori* approach aimed at deciphering group-specific DNA fragments. First, a genome-wide association study using an alignment-free method was used to detect differential *kmers* [5]. Second, the accessory genome was explored to search for group-specific orthologous genes [6]. III – An additional strategy allowed to search plasmids to characterize P or NP groups.

The workflow was tested with 59 *Erwinia*'s genomes that were manually annotated as pathogenic (P) and non-pathogenic (NP) based on bibliography. After processing, simple and multiple binary logistic regressions were applied to all results obtained by each strategy. From the 231 genes detected by the PREDIPATH database, and the 24 secondary metabolite clusters distributed along the genomes, five and nine of them, respectively, were able to characterize and predict the potential bacterial pathogenicity. The second strategy detected 512 *longmers* (concatenation of successive *k-mers*) with significant differential distribution between P and NP. Fifty-one *longmers* were detected only in NP genomes, and 12 of them present in 100% of NP genomes. Exclusive plasmids were detected in both groups.

The identification of specific markers, based on the manual curation of metadata provides important evidences to detect convergent evolution of polyphyletic groups and horizontal transfer in P and NP organisms. This workflow will allow the creation of exclusive datasets of markers that could be used as predictors to diagnostic the potential pathogenicity of plant associated bacteria.

Acknowledgements

The research leading to these results has received funding from the People Programme - Marie Curie Actions of the European Union's Seventh Framework Programme (FP7/2007-2013) under REA grant agreement n. PCOFUND-GA-2013-609102, through the PRESTIGE programme coordinated by Campus France, and from University Bretagne Loire, Programme d'Attractivité Post-Doctorale. It was conducted in the framework of the regional program "Objectif Végétal, Research, Education and Innovation in Pays de la Loire", supported by the French Region Pays de la Loire and Angers Loire Métropole.

References

[1] Chen LH, Yang J, Yu J, Yao ZJ, Sun LL, Shen Y and Jin Q, 2005. VFDB: a reference database for bacterial virulence factors. *Nucleic Acids Res*. 36 (Database issue):D539-D542.

[2] Jia B, Raphenya AR, Alcock B, et al. CARD 2017: expansion and model-centric curation of the comprehensive antibiotic resistance database. *Nucleic Acids Res*. 2016;45(D1):D566-D573.

[3] Pal, C., Bengtsson-Palme, J., Rensing, C., Kristiansson, E., Larsson, DGJ. (2014) BacMet: antibacterial biocide and metal resistance genes database, *Nucleic Acids Res.*, 42, D737-D743.

[4] Blin K, Wolf T, Chevrette MG, et al. antiSMASH 4.0-improvements in chemistry prediction and gene cluster boundary identification. *Nucleic Acids Res*. 2017;45(W1):W36-W41.

[5] Jaillard M, Lima L, Tournoud M, Mahé P, *et al*. (2018) A fast and agnostic method for bacterial genome-wide association studies: Bridging the gap between kmers and genetic events. *PLoS Genetics* 14(11): e1007758.

[6] Contreras-Moreira B, Vinuesa P. GET_HOMOLOGUES, a versatile software package for scalable and robust microbial pangenome analysis. *Appl Environ Microbiol*. 2013;79(24):7696-701.

PrivAS: a tool to perform Privacy-Preserving Association Studies

Thomas E. Ludwig[1,2], Reda Bellafqira[3], David Niyitegeka[3], Daniel Salas[4], Isabelle Perseil[4], Gouenou Coatrieux[3] and Emmanuelle Génin[1]
[1] Inserm, Univ Brest, EFS, UMR 1078 GGB, F-29200 Brest, France
[2] CHRU Brest, F-29200 Brest, France
[3] Unité INSERM UMR 1101 LaTIM, IMT Atlantique, Brest, France
[4] INSERM DSI - CISI, F-75013, Paris, France

Corresponding author: thomas.ludwig@inserm.fr

Abstract *In this paper we present PrivAS, a tool to perform Genome-Wide Association studies (GWAS) using the Weighted-Sum Statistic (WSS) algorithm in a Privacy-Preserving environment. The underlying scenario takes into account three interacting parties: (1) a Client, e.g. a genomic research unit, wanting to measure the association between an observed phenotype and regions of the genome; (2) a Reference Panel Provider (RPP) possessing genetic data for a Reference Panel, e.g. a priori healthy individuals of a carefully selected ancestry and (3) a Third-Party Server (TPS) with large computational capacities. Our tool and its underlying implementation preserve both state-of-the-art performances and Privacy for all parties. Indeed, through a series of hashing and encryption mechanisms, we can assure that no genetic data from neither the Client nor the RPP are visible by the other parties involved. Furthermore, only the Client is able to view a decrypted version of the WSS results.*

Keywords GWAS, WSS, Association Test, Privacy, Secure collaboration

1 Introduction

When studying a pathology, Genome-Wide Association Studies (GWAS) are very powerful to link phenotypes to certain regions of the genomes. Historically, GWAS were conducted on genotyping data, targeting common known SNPs. With the boom of Next-Generation Sequencing (NGS) we now have the ability to perform GWAS on rare variants [1]. The standard approach is the *case-control* setup, which compares two large groups of individuals: one *case* group presenting a particular phenotype (e.g. affected by a pathology) and one *control* group supposed healthy (at least in regard of the studied disease). Variants distribution between the two groups over all genes allows to highlight genes most probably linked to the phenotype. The cost of sequencing, which produces the type of data needed as input for these studies, has dramatically decreased over the years, nonetheless, with a limited budget, some research units (henceforth called Clients) prefer to sequence only affected individuals (so as to have a bigger panel characterising the studied pathology).

The Client can then try to use publicly available data from reference panel to test the association between the genomic regions (genes) and the pathology. Alas, more often than not, those data are aggregated data, that only show the frequencies of the variants among the individuals of this panel and not precise genotyping (EVS[2], ExAC[3], GnomAD[3],...), and tests relying on this type of data [4] are not as powerful. Indeed, knowing the frequency of each variants of a given gene within a panel doesn't allow to infer how many variants a given individual of the panel is carrying for this gene. When studying rare variants associated to a disease, it is essential to know if case samples are bearing more variants than control samples for a gene. To abate this problem, some Reference Panel Providers (RPP) such a FrEx [5] propose to perform association test (χ^2, CAST [6]) using the number of variants par sample for a selected gene, but more powerful methods are ineligible with this type of data.

When trying to perform a powerful association study on their data, Clients need to pool their genotyping data with RPP data that aren't aggregated either. This can prove troublesome as the genotyping data needed for these tests, such as the Weighted-Sum Statistic (WSS) test, are not easy to share

between the Client and the RPP. Indeed, they can be protected by various agreements and legislations and it can be difficult for the Client or the RPP to provide non-aggregated data to the other party without breaching these agreements. Furthermore, even when the protection of the privacy of the sequenced individuals is not problematic, research organisms are often shy when it comes to sharing their data with other research groups. This is because those groups might be seen as competitors and an entity, who has paid a lot of money to produce the data, doesn't want others to be able to fully exploit them on their own.

Here we present PrivAS, a tool that allows a Client to perform a WSS association study against the data from an RPP, without any parties being able to see the data from the other.

2 Algorithm

2.1 Canonical Notation of Variants

Most association tests compare genotypes of *case* and *control* individuals that might come from different sources. Classically, variants are extracted from VCF files [7] and are defined by a combination of chromosome, position, reference allele and alternate allele describing the genomic coordinates of the variant and the alteration it induces on the reference genome. When working with this notation over several sources, some multiallelic variants might be described in multiple ways. It is then sometimes laborious to check if two variants, from different VCF files, are indeed the same. To solve this problem, we propose to use a canonical notation, which uniquely describes each variant. Under this notation the consequence of the variant on the reference genome is directly described as `chromosome:start+length:sequence` (Fig.1).

excerpt from the first VCF File				canonical notation
CHROM	POS	REF	ALT	
X	12	GAC	G, GTC	X:13+2:- X:13+1:T

excerpt from the second VCF File				canonical notation
CHROM	POS	REF	ALT	
X	13	A	T	X:13+1:T

Fig. 1. The canonical notation unambiguously describes the effect of a variant over the reference genome. When comparing the variants such as they are described in both VCF files it is not immediately evident that the second alternate from the first VCF is in fact the same variant as the one in the second VCF file. Canonical notation provides a solution to this problem.

2.2 Weighted-Sum Statistic (WSS)

WSS is a widely used GWAS test that measures the association of a phenotype to a gene by comparing the number of genotypes in the gene's variants on a sample of *affected* and *unaffected* individuals [8].

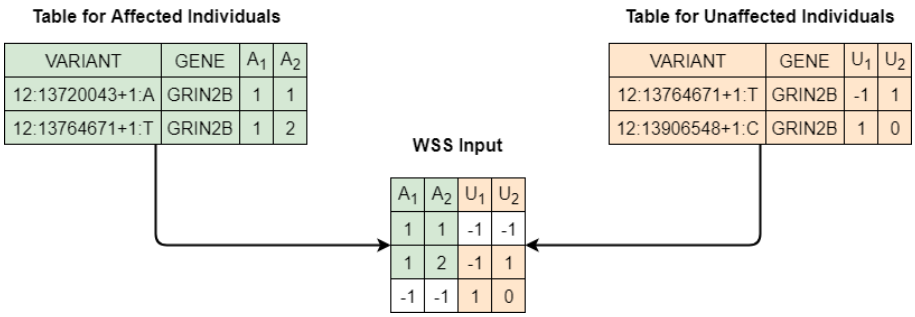


Fig. 2. Joining of case and control tables produces a suitable input for the WSS algorithm (example for 3 variants from a single gene, with 2 affected and 2 unaffected individuals).

The WSS algorithm uses two tables as input, which list variants and their associated genotypes for *affected* and *unaffected* individuals. Each row of the tables pertains to a variant. The first two columns describe the variant (in canonical notation) and the impacted gene (a variant impacting several genes

has multiple entries in the table), while the rest of the columns indicates the genotype of the variant for each individual. The genotypes are numerically coded as the number (0, 1 or 2) of variant alleles carried by a given individual (the value -1 denotes missing data for this genotype).

The preparation step of the algorithm merges these two tables, joining them on the first two columns and concatenating the genotype columns, while storing the affected/unaffected status of each individual. The resulting table is then split, grouping variants impacting the same gene. A table per gene is thus produced, from which the describing columns are stripped (Fig.2). The algorithm from [8] is then applied. It evaluates the hypothesis of the association of a gene to a phenotype and produces a p_{value} measuring the probability of finding the observed (or more extreme) results, when the null hypothesis of the underlying model is true.

2.3 Secure data exchange and computation

In our implementation of the secure WSS, three parties are involved: (1) the Client that possesses data for individuals presenting the studied phenotype; (2) the RPP that has data for unaffected individuals and (3) the Third-Party Server (TPS) that will do the actual computation. In order to allow these parties to work together without compromising the privacy of the data, encryption and hashing mechanisms will be implemented. The TPS will execute the WSS algorithm using the input tables described in the previous section where the variant (in its canonical notation) and the name of the gene affected by this variant have been hashed. This hashing is done using the SHA256 algorithm [9] initialised with a key K_{hash} shared by the Client and the RPP but unknown to the TPS. When hashing the gene names, the Client will keep a dictionary that will later allow to retrieve a gene name from a hash. As the Client doesn't have direct access to the TPS, the Client data will transit through the RPP server. Since the RPP knows K_{hash}, it is able to intercept a lot of the Client's data. To protect these data, they are encrypted using the AES algorithm [10] with a key K_{AES} generated by the Client. As the TPS needs to be able to decipher the Client's data, it also needs to know K_{AES}. So, the Client sends K_{AES} to the TPS via the RTT, protecting the key from RTT by using an RSA encryption [11]. The Client uses the public RSA key from the TPS K_p^{TPS} (that is publicly known and certified) and encrypts K_{AES} with it. Later the TPS uses its secret RSA key K_s^{TPS} (only known by the TPS) to decrypt the message. Once all computations are done, the TPS sends the results (that contain hashed gene names and their estimated p_{value}) to the Client via the RPP. To protect those results, they are encrypted used the AES key K_{AES} from the Client. The reason the Client provides an AES key and not a RSA key pair is that messages encrypted through RSA are much larger, and since genetic data are already very large it would be an unnecessary overhead. Finally, the Client uses its prebuilt dictionary to unhash the gene names.
Here is the step-by-step workflow of our secure solution (Fig.3).

1. Client gets RSA K_p^{TPS} from TPS
2. Client gets the session's unique SHA256 hash key K_{hash} from RPP
3. Client and RPP use K_{hash} to hash variants and gene names, producing WSS_{Client} and WSS_{RPP}, Client builds hash dictionary
4. Client generates a unique AES key K_{AES}
5. Client uses K_{AES} to encrypt WSS_{Client} and sends $E^{K_{AES}}(WSS_{Client})$ to RPP
6. Client uses K_p^{TPS} to encrypt K_{AES} and sends $E^{K_p^{TPS}}(K_{AES})$ to RPP
7. RPP sends WSS_{RPP}, $E^{K_{AES}}(WSS_{Client})$ and $E^{K_p^{TPS}}(K_{AES})$ to TPS
8. TPS uses RSA K_s^{TPS} to retrieve K_{AES}
9. TPS uses K_{AES} to retrieve WSS_{Client}
10. TPS performs WSS association tests for each $hash^{K_{hash}}(gene)$
11. TPS produces a $hashed.result.table$, listing each $hash^{K_{hash}}(gene)$ to its WSS p_{value}
12. TPS uses K_{AES} to encrypt $hashed.result.table$ and sends $E^{K_{AES}}(hashed.result.table)$ to RPP
13. RPP sends $E^{K_{AES}}(hashed.result.table)$ to Client

14. Client uses K_{AES} to retrieve *hashed.result.table*
15. Client uses hash dictionary on each $hash^{K_{hash}}(gene)$ to get *result.table*

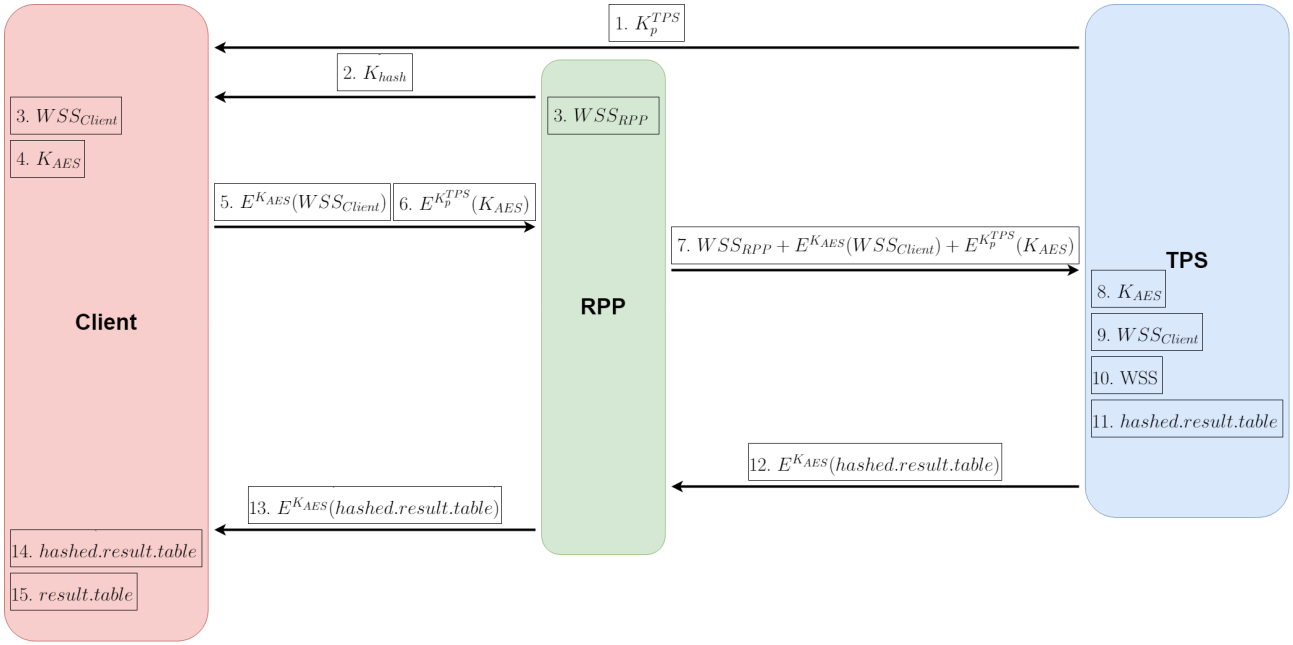


Fig. 3. The workflow of our privacy-preserving association tester.

3 Implementation

PrivAS is available as a Java Archive (jar) that can be run on the Client computer and both on RPP and TPS servers. By providing the URL of the RPP, the Client is able to launch a GWAS against its data. The transfer of data between the Client and the RPP, and between the RPP and the TPS strictly follow the algorithm depicted in the previous section, insuring the complete preservation of the privacy of all genetic data. The Client has full access to all encryption and hashing keys used for the data transfers.

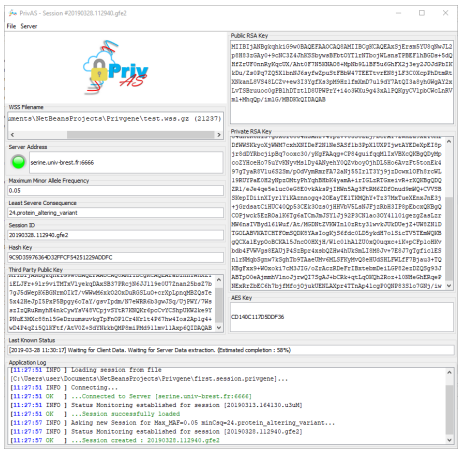


Fig. 4. The Client GUI for PrivAS displays all the information pertaining to the current association study session, such as frequency and variant consequence thresholds as well as cryptographic keys.

For the Client, PrivAS presents itself as user-friendly GUI, allowing to choose the data, set the various criteria for the study, view the keys, follow the progression of the computation and finally easily retrieve, visualise and sort the results (Fig.4). In order to fine tune the association test, the user chooses which variants will be selected based on the maximum allele frequency allowed (in GnomAD) and the least severe consequence of the variant on genes. The consequences and frequencies are

extracted from the vep annotations [12] previously added to the input VCF files. The selected criteria will be shared between the Client and the RPP, insuring a homogeneous variant selection. A given variant can be selected several times, if it affects multiple genes with consequences above the selected threshold. The Client-side of PrivAS relies on the concept of *session* to interact with the RPP server. This allows to save/reload GWAS sessions, and restore them after exiting the client. Indeed, if the study is set to run over every gene, the computation can be quite long and it is expected of the client to reconnect to the session periodical to check the progress and once all computations are over to retrieve the results.
For the RPP, PrivAS is a simple command that uses a short configuration file to act as a server waiting for Clients and to interact with the TPS.
For the TPS, PrivAS is a command that rely on a configuration file to process RPP requests.

4 Results and Discussion

4.1 Availability

The binary of PrivAS can be downloaded from `http://lysine.univ-brest.fr/privas/` and a running RPP server can be reached at `serine.univ-brest.fr` port 6666. This server uses the data from FrEx [13]. FrEx data result from the exome-wide sequencing of 574 *a priori* healthy individuals from 6 regions of France. A set of example Client data are also available to test PrivAS. The chosen TPS will be the DATARMOR supercomputer from the IFREMER (French Research Institute for Exploitation of the Sea), a research organisation independent from ours, thus insuring no collusion between the RPP and the TPS.

4.2 Robustness

The following analysis considers the semi-honest adversary model. Here each party can be considered as *honest but curious*. The corrupted parties will follow the protocols, however, the adversary is able to view the internal states (input/output and intermediate results) of every corrupted parties. We also suppose that there is no collusion between any two parties involved in the protocol. Here we analyse the possibility for one of the parties to recover another party's genotypic or haplotipic data.

Protection against the Client: The RPP's data are never, under any form, in the Client's hand. Only the client's data and the association results are available to the Client.
Protection against the RPP: Client's data and test results transit through the RPP after being encrypted via the AES cryptosystem, the security of which has been demonstrated in [10]. The key to the AES encryption also transits through the RPP and is protected by an RSA encryption that only allows the TPS to retrieve the AES key [11].
Protection against the TPS: both Client and RPP's data are at the TPS's disposal. This is indispensable as the TPS is the one in charge of the computation. However, sensible data, such as variants positions and gene names, are protected by the secure hash function SHA256. Its security has been investigated in [9]. It is not possible for the TPS to retrieve the original sensitive attribute values from their hash values. The fact that RPP will send several times its data to TPS for different studies is not a problem. Using a new secret hash key when computing SHA256 hashes makes this procedure semantically secure.

4.3 Evolution

Beyond allowing to perform secure WSS association tests, our framework can easily be extended to including other algorithms that rely on the same type of data. For example, CAST, SKAT and SKAT-O are suitable association tests that could be implemented. An evolution could also be to allow multiple Clients to pool their data together without disclosing them to each other. This could be to either have a bigger *case group* or to perform analysis such as clustering or principal component analyses over shared data. In fact, PrivAS could be extended to handle any kind of algorithm that rely on genotypic data and don't require to known genomic position.

5 Conclusion

Here we have presented PrivAS, a tool to perform Privacy-preserving WSS association studies between data from two research organisms. It takes advantage of encryption in order to secure communications between involved parties, and of a secure hash function to preserve the confidentiality of sensitive genotypic and haplotipic data. Our tool can be extended to other GWAS algorithms and to non-GWAS analyses. PrivAS, with working examples included, can be downloaded and a RPP server allowing access to the FrEx data is available.

Acknowledgements

This work is supported by *Labex CominLabs* in collaboration with *Labex Genmed*.

References

[1] Aude Saint Pierre and Emmanuelle Génin. How important are rare variants in common disease? *Briefings in Functional Genomics*, 13(5):353–361, September 2014.

[2] Exome Variant Server. Nhlbi go exome sequencing project (esp), seattle, wa. *http://evs.gs.washington.edu/EVS/*.

[3] Monkol Lek, Konrad J. Karczewski, Eric V. Minikel, Kaitlin E. Samocha, Eric Banks, Timothy Fennell, Anne H. O'Donnell-Luria, James S. Ware, Andrew J. Hill, Beryl B. Cummings, Taru Tukiainen, Daniel P. Birnbaum, Jack A. Kosmicki, Laramie E. Duncan, Karol Estrada, Fengmei Zhao, James Zou, Emma Pierce-Hoffman, Joanne Berghout, David N. Cooper, Nicole Deflaux, Mark DePristo, Ron Do, Jason Flannick, Menachem Fromer, Laura Gauthier, Jackie Goldstein, Namrata Gupta, Daniel Howrigan, Adam Kiezun, Mitja I. Kurki, Ami Levy Moonshine, Pradeep Natarajan, Lorena Orozco, Gina M. Peloso, Ryan Poplin, Manuel A. Rivas, Valentin Ruano-Rubio, Samuel A. Rose, Douglas M. Ruderfer, Khalid Shakir, Peter D. Stenson, Christine Stevens, Brett P. Thomas, Grace Tiao, Maria T. Tusie-Luna, Ben Weisburd, Hong-Hee Won, Dongmei Yu, David M. Altshuler, Diego Ardissino, Michael Boehnke, John Danesh, Stacey Donnelly, Roberto Elosua, Jose C. Florez, Stacey B. Gabriel, Gad Getz, Stephen J. Glatt, Christina M. Hultman, Sekar Kathiresan, Markku Laakso, Steven McCarroll, Mark I. McCarthy, Dermot McGovern, Ruth McPherson, Benjamin M. Neale, Aarno Palotie, Shaun M. Purcell, Danish Saleheen, Jeremiah M. Scharf, Pamela Sklar, Patrick F. Sullivan, Jaakko Tuomilehto, Ming T. Tsuang, Hugh C. Watkins, James G. Wilson, Mark J. Daly, Daniel G. MacArthur, and Exome Aggregation Consortium. Analysis of protein-coding genetic variation in 60,706 humans. *Nature*, 536(7616):285–291, August 2016.

[4] Michael H. Guo, Lacey Plummer, Yee-Ming Chan, Joel N. Hirschhorn, and Margaret F. Lippincott. Burden Testing of Rare Variants Identified through Exome Sequencing via Publicly Available Control Data. *The American Journal of Human Genetics*, 103(4):522–534, October 2018.

[5] Thomas E. Ludwig, Emmanuelle Génin, and FREX Consortium. Rare variants association test against the frex panel. *http://lysine.univ-brest.fr/FrExAC/association*.

[6] Stephan Morgenthaler and William G. Thilly. A strategy to discover genes that carry multi-allelic or mono-allelic risk for common diseases: a cohort allelic sums test (CAST). *Mutation Research*, 615(1-2):28–56, February 2007.

[7] Petr Danecek, Adam Auton, Goncalo Abecasis, Cornelis A. Albers, Eric Banks, Mark A. DePristo, Robert E. Handsaker, Gerton Lunter, Gabor T. Marth, Stephen T. Sherry, Gilean McVean, and Richard Durbin. The variant call format and VCFtools. *Bioinformatics*, 27(15):2156–2158, August 2011.

[8] Bo Eskerod Madsen and Sharon R. Browning. A groupwise association test for rare mutations using a weighted sum statistic. *PLoS genetics*, 5(2):e1000384, February 2009.

[9] S. Bakhtiari, R. Safavi-Naini, and J. Pieprzyk. Cryptographic Hash Functions: A Survey. Technical report, 1995.

[10] Joan Daemen and Vincent Rijmen. *The Design of Rijndael: AES - The Advanced Encryption Standard*. Information Security and Cryptography. Springer-Verlag, Berlin Heidelberg, 2002.

[11] R. L. Rivest, A. Shamir, and L. Adleman. A method for obtaining digital signatures and public-key cryptosystems. *Communications of the ACM*, 1978.

[12] William McLaren, Laurent Gil, Sarah E. Hunt, Harpreet Singh Riat, Graham R. S. Ritchie, Anja Thormann, Paul Flicek, and Fiona Cunningham. The Ensembl Variant Effect Predictor. *Genome Biology*, 17(1):122, June 2016.

[13] Redon R. Deleuze J.-F. Campion D. Lambert J.-C. Dartigues J.-F. Genin, E. and FREX Consortium. The french exome (frex) project : a population-based panel of exomes to help filter out common local variants. *International Genetic Epidemiology Society*, 2017.

ProteoCardis: an intestinal metaproteome-wide association study of coronary artery disease

Ariane BASSIGNANI[1,2,3], Magali BERLAND[1], Sandra PLANCADE[2], Proteocardis Consortium, and Catherine JUSTE[3]

[1] MetaGenoPolis, INRA, Domaine de Vilvert, 78350 Jouy-en-Josas, France
[2] MaIAGE, INRA, Domaine de Vilvert, 78350 Jouy-en-Josas, France
[3] Micalis, INRA, Domaine de Vilvert, 78350 Jouy-en-Josas, France

Corresponding Author: ariane.bassignani@inra.fr

In the MetaCardis FP7 framework, the gut metagenomes of more than 2000 patients at different stages of their cardiometabolic disease have been sequenced. ProteoCardis moves beyond the functional potential addressed by metagenomics, getting closer to the real functional features of the gut microbiome brought by metaproteomics that can predict aggravation of the cardiovascular risk. The metaproteomics remains a recent field, the profiling of the metaproteomes thus requires an evaluation and adaptation of bioinformatics tools originally developed for proteomics.

We extracted the gut microbiota of 138 declared coronary artery disease (CAD) patients and 50 controls, and fractionated each lysate into its cytosolic and envelope-enriched fraction that were analyzed separately on an Orbitrap Fusion™ Lumos™ Tribrid™ mass spectrometer, giving 188 individual cytosolic and as many envelope-associated datasets, plus 2x56 technical replicates. We have benchmarked several approaches of mass spectral interpretation, the samples being analyzed by tandem mass spectrometry (LC-MS/MS). In particular, we compared the reference databases to be used, as well as the identification workflows.

The optimization of these methods allows us to identify nearly 300,000 peptides and 57,000 proteins for the whole cytosolic dataset, and still more for the envelope-enriched dataset, each including 236 LC-MS/MS runs from CAD patients, controls, and replicates. The original methodology of independently analyzing cytosols and envelopes allows to identify many envelope proteins that are difficult to isolate when the cells are analyzed undivided. Statistical analyses allows us to discover metaproteomic variables that differ between groups of patients with various CAD and controls.

In addition to define new standards and practice for quantitative metaproteomics, ProteoCardis will connect gut metaproteomics data with CAD phenotypes or evolution of the health of high-risk patients to discover if the outcome in these patients can be related to particular functionalities of their microbiota.

PSH, une fonction de hachage issue du domaine du traitement d'images, permettant l'indexation et la comparaison de séquences ADN

Jocelyn DE GOËR DE HERVE[1-2], Myoung-Ah KANG[2] et Engelbert MEPHU-NGUIFO[2]

[1] UMR EPIA - INRA, VetAgro Sup, 63122, Saint Genès Champanelle, France

[2] UMR LIMOS - Université Clermont Auvergne, CNRS, 63178, Aubière Country

Corresponding Author: jocelyn.degoer@inra.fr

Abstract *L'accroissement constant des capacités de séquençage de l'ADN entraîne l'émergence de nouveaux questionnements biologiques. Le stockage et le traitement de cette masse d'informations restent des enjeux majeurs pour les années à venir. Durant le processus d'analyse des données génomiques, la recherche de séquences exactes ou proches, au travers de bases de données de génomes de références, est une tâche incontournable. Elle est notamment nécessaire dans les phases d'assemblage, d'alignement de séquences et plus généralement pour identifier la séquence de référence la plus proche d'une séquence requête. Ces tâches sont notamment essentielles dans le cadre d'étude en Biologie Évolutive, en Phylogénie ou en Métagénomique.*

Nous présentons la fonction de hachage PSH (Perceptual Sequence hashing), permettant l'indexation de séquences ADN. La fonction PSH exploite des concepts de hachage perceptuel utilisés habituellement pour indexer et comparer des images numériques, que nous avons adapté à la problématique de comparaison des séquences ADN. Outre une diminution importante des données indexées par rapport aux séquences fournies en entrée, PSH a la particularité de conserver la propriété de comparabilité entre deux clés de hachages. À partir de deux séquences ADN proches, PSH renverra des clés de hachage également proches et ainsi comparables.

Cet article présente les différentes étapes de calcul de la fonction PSH à partir d'une séquence ADN, puis la fonction de mesure de correspondance entre deux clés de hachage. Par la suite, il présente une expérimentation réalisée à partir d'un panel de séquences de références.

Les résultats obtenus en matière de sensibilité et de diminution des données, démontrent un réel intérêt quant à l'utilisation de ce type de méthodes au sein de pipelines bio-informatique. Ce travail se place dans un contexte d'accroissement des volumes de données génomiques, où l'enjeu est de concevoir des algorithmes permettant d'identifier rapidement les génomes de références les plus proches d'une séquence requête. Le but étant d'effectuer un prétraitement rapide, permettant de ne conserver que des séquences pertinentes et d'utiliser par la suite des méthodes plus classiques en bio-informatique.

Keywords Fonction de hachage, indexation, comparaison, séquence ADN, algorithme, base de données, pipeline, recherche exacte, recherche approchée, structure de données, table de hachage.

1. Introduction

L'évolution constante des techniques de séquençage de l'ADN, entraîne la production de plus en plus massive de données génomiques, pour un coût de plus en plus bas. L'apparition de séquenceurs toujours plus modulables et portatifs [1], qui permettent la lecture de fragments d'ADN, de plus en plus long, va conduire à une démocratisation certaine de ces outils. Outre l'émergence de nouveaux questionnements biologiques [2],

le stockage et les besoins d'analyse rapide de cette masse d'informations est un enjeu majeur pour les années à venir. Après les étapes de séquençage, l'une des premières étapes de leur analyse, est la caractérisation des séquences ADN obtenues, afin de déterminer le génome de référence le plus proche, en utilisant des bases de données de références. La recherche de fragments de séquences ADN, au travers de bases de données est aussi utilisée pour les tâches d'assemblage de plusieurs séquences à partir d'une référence [3], pour étudier les mutations ou pour déterminer une sous-séquence commune entre plusieurs séquences. Afin de répondre à ces problématiques, de nombreuses méthodes en bio-informatique ont été proposées, utilisant des techniques de comparaison de textes [4], d'alignement local [5], d'indexation de k-mers [6], d'utilisation d'arbres des suffixes [7] comme structure de données. Bien que ces méthodes soient reconnues comme étant très efficaces, leur utilisation peut être limitée de par leur complexité algorithmique, ce qui peut avoir pour conséquence de rendre difficile le changement d'échelle en termes de taille des données dans les années à venir. La littérature décrit aussi l'utilisation de table de hachage [8], permettant la recherche exacte et l'alignement de séquences ADN. Le processus d'indexation, nécessite le découpage en sous-séquences (k-mer), des séquences à indexer et pour chacun d'entre eux de conserver leurs positions sur la séquence dont ils sont issus.

Compte tenu de l'accroissement des données à analyser, l'un des enjeux est de créer des algorithmes permettant d'identifier rapidement les génomes de référence les plus proches, d'une séquence nouvellement séquencée. Ceci afin d'effectuer un rapide prétraitement permettant de conserver uniquement des séquences de références pertinentes pour qu'elles soient par la suite analysées par des outils plus spécifiques aux questionnements biologiques.

Cet article a pour objectif de présenter la fonction de hachage PSH (Perceptual Sequence Hashing). PSH, est une fonction de hachage perceptuel, dérivée de fonctions de hachage permettant l'indexation d'images numériques. Elle a été adaptée de façon à pouvoir répondre aux caractéristiques intrinsèques des séquences ADN. L'objectif a consisté à développer une fonction de hachage permettant l'indexation de séquences ADN, de façon à entrainer une diminution importante de données tout en conservant la possibilité de les comparer. Le processus de calcul des clés de hachage binaires correspondants aux séquences ADN, consiste tout d'abord à les convertir sous la forme de matrice de pixels afin d'en extraire des caractéristiques visuelles à l'aide d'une fonction TCD [9] exploitant le domaine fréquentiel. Le processus de comparaison de deux clés de hachage est effectué via une Distance de Hamming, qui permet d'obtenir une notion d'homologie entre deux séquences ADN.

1.1. Motivations

Il existe certaines similitudes entre les données de type images et les données génomiques. Tout d'abord, leurs modes de production n'ont cessé d'évoluer depuis plusieurs décennies, ce qui soulève de nombreuses problématiques en termes de stockage, mais aussi et surtout en ce qui concerne leur analyse. En effet, les images numériques et les données génomiques sont toutes issues d'un processus d'acquisition et subissent une transformation au format numérique, avec pour l'une, l'acquisition de la réfraction de la lumière et pour l'autre la lecture d'un enchaînement de molécules. Une fois numérisées, ceci a pour conséquence la production de documents numériques, composés de sous unités. Ainsi, une image numérique est un signal discret pouvant être composé de quelques dizaines à plusieurs millions de sous-unités appelées pixels. De même, une séquence ADN est aussi un signal discret pouvant être aussi composé de quelques dizaines à plusieurs millions de sous-unités appelées nucléotides. Dès lors, les processus d'analyse de base sont semblables, puisqu'ils nécessitent de pouvoir comparer tout ou partie des images ou des séquences ADN. De plus les phases de comparaison doivent aussi prendre en compte des facteurs de dégradation, notamment le redimensionnement ou la compression avec perte, pour les images numériques et les erreurs de séquençage ou les mutations pour les séquences ADN. Cependant l'une des différences majeures réside dans le fait que les images aient deux dimensions, alors que les séquences ADN n'en possèdent qu'une.

Durant ce travail, nous sommes partis du postulat émis par MC. Saldías et al. [10], que les séquences ADN pouvaient être encodées sous formes d'images numériques, au sein de matrices de pixels à deux dimensions. Intuitivement, il aurait sans doute été plus naturel d'envisager des approches basées sur des méthodes de comparaison de signaux audio à une dimension. Cependant, Y. Ke et al. [11] ont décrit une technique d'identification de musique via une méthode d'analyse d'images, où le signal audio, un signal à une dimension (1D), est converti en plusieurs images à deux dimensions (2D). Cette méthode qui donne des résultats très satisfaisants, démontre l'intérêt d'effectuer une conversion d'un signal 1D en un signal 2D, afin de ne plus se

baser sur les caractéristiques d'un signal audio mais sur sa représentation sous forme d'images. Le passage à deux dimensions apportant plus de diversité, ce qui est nécessaire au calcul des points d'intérêts qui servent de bases au calcul des clés de hachage.

Ceci est d'autant plus vrai pour les séquences ADN, puisqu'elles sont constituées de sous-éléments ne pouvant prendre que quatre états différents (*A, T, C, G*) correspondants aux quatre types de nucléotides (Adénine, Thymine, Cytosine, Guanine). Nous avons donc opté pour une conversion des séquences ADN sous la forme de matrices de pixels, en attribuant des tailles arbitraires à chacune des deux dimensions. Ce processus est totalement réversible puisque lors de la phase de conversion, le nucléotide suivant le dernier d'une ligne se trouve être le premier de la ligne suivante.

1.2. Organisation

Nous présentons de façon détaillée, les différentes étapes de calcul de la fonction PSH ainsi que la méthode de mesure de correspondance permettant de comparer les clés de hachage. Par la suite nous présentons une des expérimentations que nous avons réalisé à partir d'un jeu de données réelles, durant le processus de validation de la fonction PSH.

2. Matériel et méthodes

La fonction PSH a été développée de façon à permettre une diminution des données, tout en conservant la possibilité de les comparer. À l'instar des fonctions de hachage perceptuel, l'objectif était de pouvoir utiliser des structures de type table de hachage pour stocker les données de séquences ADN indexées via la fonction PSH. Ce type de structure permet par la suite de pouvoir interroger les données, en exploitant la faible complexité algorithmique d'interrogation des tables de hachage.

La fonction PSH accepte en entrée des séquences de taille *W* et renvoie des clés de hachage de taille *W'*. L'objectif est de calculer, pour une séquence *S* de taille *W*, une emprunte représentative, au format binaire de taille fixe *W'*, appelée clé de hachage *H*. Une clé de hachage est directement représentative de la séquence *S* et sa taille *W'* est toujours inférieure à *W*. À l'instar des fonctions de hachage, la fonction PSH n'est pas réversible, ce qui signifie qu'à partir d'une clé de hachage il n'est pas possible de recalculer la séquence d'origine dont elle est issue.

Le principe général de la fonction PSH consiste tout d'abord à extraire des caractéristiques représentatives des matrices de pixels générées à partir des séquences ADN dont on souhaite calculer la clé de hachage. Une fois ces caractéristiques extraites, un algorithme de compression d'images est appliqué avec un fort taux de compression. Ceci a pour effet, d'entraîner une forte dégradation de la matrice contenant les caractéristiques de l'image. Cependant, même fortement dégradée la matrice de pixels résultant de la compression, reste toujours représentative de l'image originale, donc de la séquence ADN dont elle est issue. Au final, les clés de hachage générées résultent d'une surcompression de l'image à l'intérieur duquel elles sont encodées.

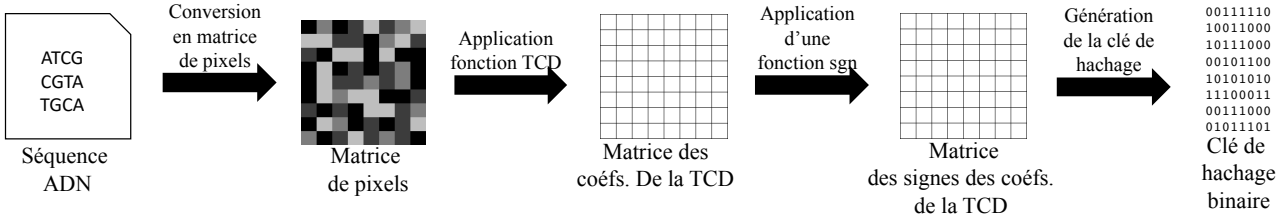


Fig 1. *Principales étapes de calcul de la fonction PSH*

2.1. Principales étapes de calcul de la fonction PSH

La première étape de la fonction PSH est la conversion des séquences ADN sous forme d'images, ou plus précisément, leur encodage au sein d'une matrice de pixels de taille *N x M* (Cf. Fig.1). Ainsi, chacun des nucléotides composant une séquence ADN devient un pixel de l'image. Les pixels sont disposés ligne par ligne en partant du coin supérieur gauche de l'image. Le nucléotide suivant directement le dernier d'une ligne, sera le premier de la ligne suivante. Il est à noter que cette méthode d'encodage n'a aucun fondement ni justification

biologique directe. Cependant, il s'agit d'une façon de réorganiser l'information, tout comme le texte d'un livre est mis en page afin d'occuper un maximum de place sur le papier où il est imprimé.

Ainsi, chaque nucléotide devient un pixel, ayant une valeur d'intensité lumineuse (L_1, L_2, L_3 et L_4), en fonction de son type (*A, T, C, G*). Les pixels sont encodés au sein d'une échelle ayant 256 valeurs possibles de niveaux de gris, sur 1 octet. Ce type d'encodage correspond au type « caractère » utilisé par le format FASTA. À ce stade, il n'y a donc aucune diminution ni dégradation de la séquence et le processus d'encodage est une opération totalement réversible.

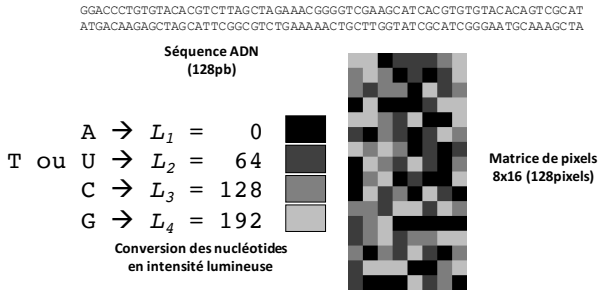


Fig 2. *Conversion d'une séquence ADN en matrice de pixels*

À partir des séquences ADN, le processus d'encodage crée des images qui se rapprochent plus d'un signal aléatoire, que d'une image ayant une véritable structure. L'objectif de cet encodage, qui reste arbitraire, est de créer une certaine diversité anthropique qui permettra par la suite l'extraction de points d'intérêts significatifs et représentatifs. À l'instar de Saldías et al. [11], nous avons opté pour une approche exploitant le domaine fréquentiel afin de calculer des points d'intérêt. En effet, la littérature décrit de nombreuses fonctions permettant le calcul de descripteurs d'images. Cependant, étant donné le caractère peu structuré des images produites à partir des séquences ADN, qui se rapprochent plus d'un signal aléatoire, que d'une véritable image structurée, les descripteurs globaux s'appuyant sur le domaine fréquentiel nous ont paru être les descripteurs les plus pertinents pour notre problématique.

2.2. Principales étapes de calcul de la fonction PSH

La seconde étape consiste à appliquer une fonction mathématique appelée Transformée en Cosinus Discrète (TCD) [9], sur la matrice de pixels. La fonction TCD permet de faire passer les données depuis le domaine spatial vers le domaine fréquentiel. Elle est réversible via sa fonction inverse (TCDi), ce qui n'entraine à ce stade aucune perte d'information. La TCD autorise simplement un changement de domaine d'étude, tout en gardant la même fonction étudiée.

Lorsque la fonction TCD est appliquée à une image, sa représentation fréquentielle est obtenue dans une matrice appelée matrice des coefficients fréquentiels. La matrice des coefficients fréquentiels a les mêmes dimensions que la matrice de pixels originale, dont elle est issue. Au sein d'une image, les fréquences représentent les variations de l'intensité des pixels. La TCD a pour caractéristique principale de regrouper les hautes fréquences dans la partie supérieure gauche de la matrice des coefficients. Les hautes fréquences portent l'information relative à la structure des images, c'est à dire, les changements d'intensité rapides qui correspondent aux contours des formes. Tandis que les basses fréquences qui correspondent aux changements d'intensité lents, portent les zones homogènes.

2.3. Seconde étape : Récupération des signes des coefficients de la matrice des coefficients

Cette étape a pour objet d'extraire la structure représentative de l'image de départ. Elle est calculé à partir de la matrice des coefficients de la TCD, à laquelle on applique une fonction sgn() afin de ne conserver que les signes des coefficients.

Cette méthode est notamment décrite par H. Stark [12] dans l'ouvrage « Image recovery : Theory and Application », sous la dénomination de « Sign-Only Synthesis » (SOS). Le concept de SOS est applicable à de multiples fonctions permettant la représentation fréquentielle d'un signal telles que la Transformée de Hadamard, la Transformée de Fourier ou la Transformée de Karhunen-Loeve. De façon générale, il est défini comme étant la Transformée Inverse de la matrice binaire des signes des coefficients.

Les signes des coefficients de la TCD, ont un rôle essentiel car ils sont toujours porteurs de l'information structurelle de l'image et ceci malgré la suppression des valeurs des coefficients fréquentiels. Cette étape n'est pas réversible, étant donné l'importante suppression des données dû à l'unique conservation des signes des coefficients.

2.4. Seconde étape : Récupération des signes des coefficients de la matrice des coefficients

Les clés de hachage sont calculées à partir de la matrice des signes des coefficients. Par conséquent, elles sont nativement au format binaire. Le nombre de coefficients binaires étant directement en relation avec le nombre de coefficients de la TCD, qui sont eux même directement issus du nombre de pixels de l'image d'origine ; la taille maximale d'une clé de hachage ne peut donc pas être supérieure à celle-ci. En fonction de la taille que l'on souhaite donner aux clés de hachage, il est possible de ne conserver qu'une partie des coefficients binaires. Dans ce cas, on notera que plus une clé de hachage a une taille réduite, plus sa probabilité de collision augmente et moins elle contient d'informations structurelles représentatives de l'image d'origine.

2.5. Seconde étape : Récupération des signes des coefficients de la matrice des coefficients

Étant donné le mode de calcul des clés de hachage, nous pouvons en déduire que la fonction PSH entraine une diminution importante des données. En effet, si la clé de hachage est formée à partir de tous les coefficients de la matrice binaire, elle encodera chaque nucléotide sur un seul bit de données. De plus, à l'instar de l'algorithme de compression JPEG, il est possible de ne garder qu'une partie des coefficients correspondants au hautes fréquences de la matrice binaire pour générer la clé de hachage. Une clé de hachage encodée sur 64bits (8 octets) à partir d'une séquence de 256pb, sera donc 32x moins volumineuse que la séquence dont elle est issue.

2.6. Comparaison des clés de hachage

Une des caractéristiques essentielles de la fonction PSH, est qu'il est possible de déterminer une distance entre deux clés de hachage. Ceci est dû à la manière dont elles sont calculées, car, malgré le fait qu'elles soient au format binaire, elles contiennent toujours une partie de l'information structurelle de la matrice de pixels (image), dont elles sont issues (propriété de représentativité). Afin de calculer la distance entre deux clés de hachage, la distance de Hamming [13] est utilisée. La Distance de Hamming est la distance de référence pour comparer deux chaines de tailles identiques. Elle exprime la somme des différences entre deux séquences de même longueur. Les séquences peuvent être des suites de nombres binaires mais aussi se composer d'éléments provenant d'autres systèmes numériques ou alphanumériques. La distance de Hamming renvoie donc un indice de distance : plus cet indice est faible et plus les séquences sont similaires. En revanche, la distance de Hamming entre deux clés de hachage, n'est pas assez précise pour permettre de déterminer le taux de mutation entre deux séquences.

2.7. Évaluation

L'expérimentation que nous avons menée avait pour objectif d'évaluer la recherche de k-mers proches mais non exactes au sein d'une table de hachage, en utilisant les propriétés de comparabilité des clés de hachage calculées par la distance de Hamming. Une base de données, de 4 séquences d'ADN de référence, a été constituée (Cf. *Tab.1*). Ces séquences ont été choisies car elles ne présentaient aucune homologie. Elles ont été indexées et stockées au sein d'une structure de données appelée PSH-DB, que nous avons développé pour les expérimentations. Ainsi, chaque séquence a été découpée en sous-séquence ou k-mers, d'une taille de 128pb, avec un décalage de 64pb (Cf. *Fig.3*). Pour indexer l'intégralité d'une séquence, dès lors que sa taille n'est pas un multiple de la taille des k-mers fixée en paramètre (128 pour cette expérimentation), il est nécessaire de définir un k-mer de fin de séquence. Ainsi, pour une séquence de taille de 530pb et des k-mers de 128, le k-mer de fin de séquence débutera à partir de la position 402 (530 – 128).

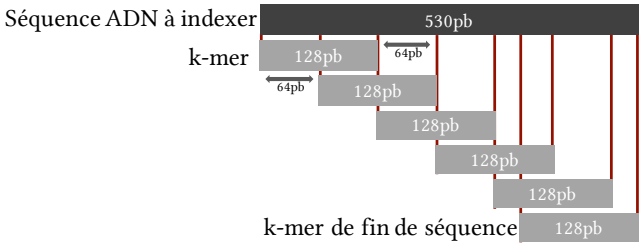


Fig 3. *Processus d'indexation d'une séquence de référence par découpage en k-mers*

Pour chaque k-mer une clé de hachage correspondante a été calculée. C'est cette clé de hachage qui a été stockée au sein de la structure PSH-DB.

L'objectif était de calculer les distances de Hamming entre les clés de hachage issues d'une séquence requête et toutes les clés présentes dans la base de données. Ceci afin de déterminer s'il était possible de fixer un seuil en fonction des paramètres utilisés et au final de retrouver la séquence de référence dont était issue une séquence requête comportant des taux de mutation.

Nom de la séquence de référence	Taille
Borrelia burgdorferi B31, complete genome	910 724 pb
Escherichia coli O157:H7 str. Sakai DNA, complete genome	5 498 450 pb
West Nile virus lineage 2, complete genome	10 962 pb
Salmonella enterica subsp complete genome	4 857 432 pb

Tab 1. *Séquences de référence utilisées pour l'expérimentation*

Durant cette expérimentation, nous nous sommes plus particulièrement focalisés sur la sensibilité de la méthode de recherche. Les temps de traitement présentés Tab. 3, ne sont donnés qu'à titre indicatif et ne peuvent pas être considérés comme représentatifs. En effet, l'expérimentation a consisté à comparer toutes les clés issues des différents k-mers des séquences requêtes avec toutes les clés de la base de données, ceci afin de pouvoir établir des statistiques sur les résultats obtenus.

2.7.1. Création du jeu de requêtes

Pour cette expérimentation, 6000 séquences requêtes, d'une taille de 1024pb ont été utilisées. Le Tab.5 présente le nombre de séquences requêtes qui ont été extraites à partir des séquences de référence et les taux de mutation qui leur ont été appliqués. Il est à noter qu'une des séquences de référence n'a pas été utilisée pour extraire des requêtes et que nous avons ajouté 1000 requêtes générées totalement aléatoirement.

Séquences de référence	Nombre de requêtes	Taux de mutation
Borrelia burgdorferi	1000	5%
Borrelia burgdorferi	1000	10%
Escherichia coli	1000	5%
Escherichia coli	1000	10%
WestNile	1000	5%
WestNile	1000	10%
Salmonella enterica	0	-
Séquences aléatoires	1000	-

Tab 2. *Séquences de référence utilisées pour la génération de requêtes pour l'expérimentation*

3. Résultats

Cette expérimentation, avait pour objectif d'évaluer la possibilité de retrouver pour chaque séquence de notre panel, les séquences de référence les plus proches, malgré des taux de mutation compris entre 5% et 10%. Les résultats illustrés par la Fig.4 montrent la distribution cumulée totale des distances de Hamming, calculées à partir des séquences requête ayant été générées depuis les séquences de référence, avec un taux de mutation de 10%. La colonne « positifs » représente le nombre de clés de hachage appartenant aux séquences requête ayant été attribuées avec succès, à la séquence de référence dont elles étaient issues. La colonne des « négatifs » représente les clés ayant été attribuées de façon erronée à une séquence de référence.

En observant les distributions des distances de Hamming, issues de ces deux colonnes, il est possible de définir un intervalle significatif (distance de Hamming <= 8), permettant de déterminer la proximité entre deux clés de hachage, donc entre deux k-mers et par extension, une zone d'homologie entre deux séquences.

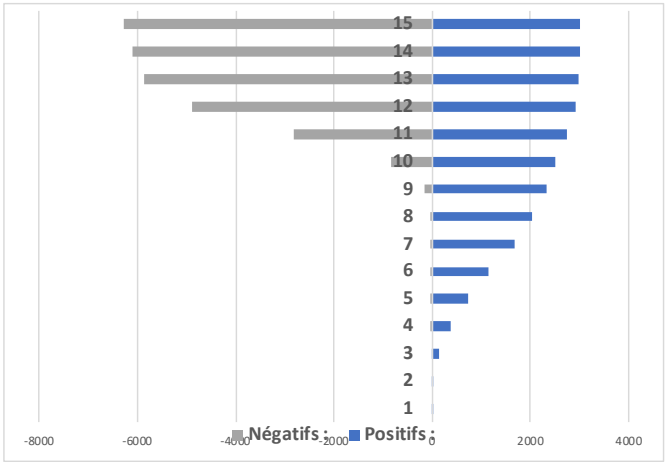


Fig 4. *Distribution du nombre de requêtes (positives ou négatives) en fonction de la Distance de Hamming avec des séquences requêtes ayant un taux de 10% de mutation*

Enfin le Tab. 3, présente la synthèse de cette expérimentation. Les résultats présentés pour les jeux de séquences requête présentant des taux de mutation de 5% et 10% montrent des résultats très encourageants avec de forts taux de détection >99% et des taux de faux positifs compris entre 2% et 3%, avec un seuil de distance de Hamming établi à 8.

Taux de mutation	Positifs (seuil <= 8)	Négatifs (seuil <= 8)	Temps d'exécution
5%	100%	2,53%	486 min
10%	99,33	2,83%	492 min

Tab 3. *Synthèse des résultats obtenus en fonction du taux de mutation*

4. Conclusion

Les méthodes d'identification de documents de type image ou audio basées sur des fonctions de hachage perceptuel sont reconnues pour leurs capacités à comparer des documents proches et pour leurs faibles complexités algorithmiques. Cependant, bien que certaines problématiques soient communes avec les données génomiques, la littérature ne décrit aucune approche concernant l'utilisation de fonctions de hachage perceptuel appliquées aux séquences ADN.

Au cours de cet article, nous avons présenté la fonction PSH. Elle a été développée avec l'idée de pouvoir proposer une fonction de hachage capable d'identifier une séquence ADN via le calcul d'une emprunte binaire de taille inférieure tout en gardant la propriété de pouvoir être comparée. La capacité de pouvoir comparer des clés de hachage, c'est à dire, de pouvoir établir une notion de distance entre deux clés est un concept qui n'avait

jamais été étudié jusqu'alors en bio-informatique. Les développements méthodologiques qui ont donné lieu à la fonction PSH sont une illustration et une application directe de ces concepts.

Cet article a tout d'abord introduit le principe général ainsi que les différentes étapes de calcul de la fonction PSH puis la méthode de comparaison des clés de hachage. Par la suite, il a présenté une des expérimentations que nous avons mené au cours de ce travail. Elle consistait en la recherche de séquences proches au sein d'une base de données de références. Nous nous sommes plus particulièrement focalisés sur l'évaluation des capacités de la fonction PSH à pouvoir être utilisée pour retrouver des séquences proches et ainsi d'en évaluer sa sensibilité.

Avec un taux de détection de 99% des séquences de référence, à partir de requêtes ayant des taux de mutation de 10%, les résultats obtenus en utilisant notre fonction PHS sont très encourageants. Ils ouvrent la possibilité d'utiliser des concepts usuels dans le domaine de l'imagerie, pour répondre à la problématique de l'indexation et la recherche de séquences ADN proches au sein de base de données de références.

Références

[1] A. B. R. McIntyre *et al.*, « Nanopore sequencing in microgravity », *Npj Microgravity*, vol. 2, nº 1, déc. 2016.

[2] M. Pop et S. L. Salzberg, « Bioinformatics challenges of new sequencing technology », *Trends Genet. TIG*, vol. 24, nº 3, p. 142-149, mars 2008.

[3] S. Schbath, V. Martin, M. Zytnicki, J. Fayolle, V. Loux, et J.-F. Gibrat, « Mapping Reads on a Genomic Sequence: An Algorithmic Overview and a Practical Comparative Analysis », *J. Comput. Biol.*, vol. 19, nº 6, p. 796-813, juin 2012.

[4] P. Kalsi, H. Peltola, et J. Tarhio, « Comparison of Exact String Matching Algorithms for Biological Sequences », in *Bioinformatics Research and Development*, vol. 13, M. Elloumi, J. Küng, M. Linial, R. F. Murphy, K. Schneider, et C. Toma, Éd. Berlin, Heidelberg: Springer Berlin Heidelberg, 2008, p. 417-426.

[5] S. F. Altschul, W. Gish, W. Miller, E. W. Myers, et D. J. Lipman, « Basic local alignment search tool », *J. Mol. Biol.*, vol. 215, nº 3, p. 403-410, oct. 1990.

[6] N. Välimäki et E. Rivals, « Scalable and Versatile k-mer Indexing for High-Throughput Sequencing Data », in *Bioinformatics Research and Applications*, vol. 7875, Z. Cai, O. Eulenstein, D. Janies, et D. Schwartz, Éd. Berlin, Heidelberg: Springer Berlin Heidelberg, 2013, p. 237-248.

[7] L. Feng, A. Jean, C. P. Leng, et L. Danbo, « Identification of DNA Signatures via Suffix Tree Construction on a Hybrid Computing System », vol. 9, nº 2, p. 10, 2012.

[8] S. Misra, A. Agrawal, W. Liao, et A. Choudhary, « Anatomy of a hash-based long read sequence mapping algorithm for next generation DNA sequencing », *Bioinforma. Oxf. Engl.*, vol. 27, nº 2, p. 189-195, janv. 2011.

[9] N. Ahmed, T. Natarajan, et K. R. Rao, « Discrete Cosine Transform », *IEEE Trans. Comput.*, vol. C-23, nº 1, p. 90-93, janv. 1974.

[10] Yan Ke, D. Hoiem, et R. Sukthankar, « Computer Vision for Music Identification », in *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, San Diego, CA, USA, 2005, vol. 1, p. 597-604.

[11] M. Curilem Saldías, F. Villarroel Sassarini, C. Muñoz Poblete, A. Vargas Vásquez, et I. Maureira Butler, « Image Correlation Method for DNA Sequence Alignment », *PLoS ONE*, vol. 7, nº 6, p. e39221, juin 2012.

[12] H. Stark, *Image Recovery Theory and Application.* Saint Louis: Elsevier Science, 2014.

[13] R. W. Hamming, « Error Detecting and Error Correcting Codes », *Bell Syst. Tech. J.*, vol. 29, nº 2, p. 147-160, avr. 1950.

R: Ecology Met A Data Language

Elie Arnaud[1], Yvan Le Bras[1]
[1] Station marine de Concarneau BP 225, 29182 Concarneau CEDEX, France
PatriNat UMS 2006 Paris, France

Corresponding author : elie.arnaud@mnhn.fr

Abstract

Ecology is a complex research field, relying on multiple methods – from physics to sociology – and addressing a high variety of questions concerning our environments. Thus, it requires a high number of descriptors and consequently produces metadata-rich information.

Such a complex task calls for complex standard and related tools to permit the handling of this rich information. This complexity is currently notably covered by the Ecology Metadata Language (EML) standard, provided thanks to the work of the Ecological Society of America [1] and the Knowledge Network for Biocomplexity (KNB) [2] which is the EML reference. Although EML covers a huge diversity of metadata, it keeps being hard to apprehend for neophytes informatics users.

In this context, our purpose is to offer a user-friendly way to use the complex EML metadata standard for the widest number of biodiversity centered use cases. We thus propose to investigate ways to facilitate 1/ the reading and understanding of EML, 2/ the EML files generation by using its specification within an automated method and 3/ an easy-to-use graphical interface dedicated to real life lab work.

The R language was chosen to perform this task as it is the commonest language in ecology. Also, the R package ‘shiny’ [3] provides methods to produce a user interface almost as flexible as HTML could render. This choice will ease the task of anyone who would contribute to the tool. Furtherly, a better information quality would permit to reach a high degree of FAIRness for the so-described studies [4].

References

1. Michener, William K., et al. Nongeospatial Metadata for the Ecological Sciences. *Ecological Applications*, vol. 7, no. 1, 1997, pp. 330–342. *JSTOR*, www.jstor.org/stable/2269427
2. Detailed information about EML on the KNB website: https://knb.ecoinformatics.org/external//emlparser/docs/index.html
3. Winston Chang, Joe Cheng, JJ Allaire, Yihui Xie and Jonathan McPherson (2018). shiny: Web Application Framework for R. R package version 1.2.0. https://CRAN.R-project.org/package=shiny
4. More information about the GO FAIR BiodiFAIRse initiative on their website : https://www.go-fair.org/implementation-networks/overview/biodifairse/

RandomRead : a sequence-read simulator program for metagenomic shotgun

Guillaume Gricourt[1], Mélissa N'Debi[1], Vanessa Demontant[1], Anais Nguyen-Goument[1], Abdelrazak Aissat[1,2,4], Paul-Louis Woerther[3] and Christophe Rodriguez[1,2,3]

[1] NGS platform AP-HP - IMRB Institute, Henri Mondor Hospital - University Paris-Est, 94000, Créteil, France

[2] Inserm U955, Henri Mondor Hospital - University Paris-Est, 94000, Créteil, France

[3] Department of Microbiology AP-HP, Henri Mondor Hospital - University Paris-Est, 94000, Créteil, France

[4] Department of Genetics, Henri Mondor Hospital - University Paris-Est, 94000, Créteil, France

Corresponding author: guillaume.gricourt@aphp.fr

Evaluation of the performances of analysis software used for Next Generation Sequencing is essential in research studies and require in medical diagnosis through the Norma NF EN ISO 15189. To satisfy these requirements, different datasets created *in silico* are necessary to use during the whole development of a new software, at the end to evaluate final performance and for new versions [1]. Theses datasets must be very similar to the real sequences produced by the experimental protocol [2], content must be perfectly controlled, and particular variations must be possible to test according to the goal of the project (deletion, insertion, sensitivity...). In our project of Shotgun Metagenomic detecting all micro-organisms in infectious disease, different software producing datasets are available (FASTQSim, GemSim, Grinder, Mason, NeSSM or pIRS) but have some limitations and is poorly adapted to our field.

We created a new software, RandomRead, which produce sequences from metagenomic shotgun protocol using Illumina sequencing. Entry parameters are one or several sequences of micro-organisms reference, the total number of sequences, the dilution level in human reference, and the size of fragmentation to simulate real experimental data. In addition, considering micro-organisms are variable and sequencing error are always produced, a mutation rate could be applied. Finally, a diversity of the reference is created by providing a VCF file containing mutations and their frequencies.

RandomRead is written in C++ with the popular library Htslib [3] and create 9800 sequence in one second. The software produced 120 samples with virus (with RNA and DNA genome), bacteria (Gram negative and Gram positive) and fungi (yeast and mold) at different level of dilutions and variability. These dataset were used with success to evaluate sensitivity, specificity, repetability... of our diagnosis software. The results allowed our laboratory to be accredited (NF ISO 15189, BM MG6 of SH INF 50 v06). It will be extend to fit with the genomic panel and exome for the next accreditation.

Keywords : NGS, Metagenomic Shotgun, NF EN ISO 15189, Diagnosis, Software

References

[1] Merly Escalona, Sara Rocha, and David Posada. A comparison of tools for the simulation of genomic next-generation sequencing data, aug 2016.

[2] Robert Schlaberg, Charles Y. Chiu, Steve Miller, Gary W. Procop, and George Weinstock. Validation of metagenomic next-generation sequencing tests for universal pathogen detection. *Archives of Pathology and Laboratory Medicine*, 141(6):776–786, jun 2017.

[3] Heng Li, Bob Handsaker, Alec Wysoker, Tim Fennell, Jue Ruan, Nils Homer, Gabor Marth, Goncalo Abecasis, and Richard Durbin. The Sequence Alignment/Map format and SAMtools. *Bioinformatics*, 25(16):2078–2079, aug 2009.

Recherche par clustering de gènes impliqués dans le syndrome PTSD

Justine POLLET[1], Jean-Baptiste WOILLARD[1,2] and Claire-Cécile BARROT[1,2]

[1] INSERM UMR1248, Université de Limoges, Centre de Biologie et de Recherche en Santé, Rue du Pr Descottes, 87025, Limoges, France

[2] CHU de Limoges, Laboratoire de Pharmacologie, toxicologie et pharmacovigilance, 2 avenue Martin Luther King, 87042, Limoges, France

Corresponding Author: claire-cecile.barrot@unilim.fr

La Ciclosporine et le Tacrolimus sont des inhibiteurs de la calcineurine couramment prescrit en tant qu'immunosuppresseur anti-rejet suite aux greffes d'organes. Ils inhibent par de mécanismes différents la calcineurine, bloquant la synthèse de l'interleukine 2 permettant l'activation lymphocytaire à l'origine de l'organisation de la réponse immunitaire contre le greffon, principalement les lymphocytes T cytotoxiques (CTL). Cependant, ces médicaments sont à marge thérapeutique étroite et administrés de façon chronique ce qui signifie que des effets indésirables peuvent apparaitre. Un des effets indésirables des inhibiteurs de la calcineurine est l'augmentation du risque d'apparition de syndrome lymphoprolifératif post-greffe, ou PTLD (Post-Transplant Lymphoproliferative Disease). Les PTLD constituent le deuxième groupe le plus fréquent de tumeurs malignes survenant après une transplantation d'organe solide, représentant la principale cause de décès et de perte de greffes liés au cancer. Une cause connue est l'implication fréquente du virus Epstein-Barr (EBV).

Nous avons effectué un séquençage d'exome (WES) puis un appel de variants sur une cohorte de 16 patients transplantés rénaux traités par inhibiteur de la calcineurine, appariés selon leur statut vis-à-vis de l'EBV, leur âge, sexe (8 cas PTLD et 8 témoins). La recherche de variants causaux s'étant avérée peu concluante, nous nous sommes orientés vers la recherche de nouveaux gènes d'intérêt. Une matrice de présence/absence des variants pour chaque groupe étudié (cas-témoins) a été construite, prenant en compte l'appariement cas/témoins pour filtrer les variants. Cela a permis la création d'une matrice de contingence variants/gènes, en pondérant les variants selon leur localisation exonique (poids 1) ou non-exonique (poids 0,1). Cette matrice a été normalisée en fonction de la taille des gènes, en faisant ainsi une matrice de 'degré' de variation. Enfin, nous avons effectué des clusterings kmeans en variant le nombre de clusters, afin d'identifier des groupes de gènes variants de manière similaire.

Un cluster 'noyau' a été identifié, dont les frontières sont mieux définies avec k=3. Il comporte cinquante-deux gènes, de trente-quatre familles de gènes différentes. Parmi les cinquante-deux, cinq sont associées à des voies métaboliques d'intérêts. H1FNT, HIST1H4C, HUS1B et PAXX sont tous les quatre impliquées dans des processus métaboliques de l'ADN : H1FNT et HIST1H4C associés au silencing, HIST1H4C, HUS1B et PAXX associée à de la réparation ADN. Ces 4 gènes pourraient être impliqués dans les processus tumoraux PTLD. ISG15 pourrait jouer un rôle dans la réponse à l'EBV du fait de son implication dans les processus de conjugaison ISG15-protein et dans la régulation négative de la réplication des génomes viraux. Enfin, les gènes restants sont liés à la perception sensorielle et à la sécrétion de vasopressine. Nous supposons qu'il s'agit d'un « bruit de fond » probablement causés par la petite taille de la cohorte utilisée.

Couplée à la plausibilité biologique, cette approche nous a permis d'identifier 5 gènes d'intérêt qui seront confirmés par des analyses ultérieures.

ReClustOR, a Re-Clustering tool using an Open-Reference method that improves OTU definition

Christophe DJEMIEL[1], Corentin JOURNAY[1], Battle KARIMI[1], Samuel DEQUIEDT[1], Walid HORRIGUE[1], Pierre-Alain MARON[1], Nicolas CHEMIDLIN-PREVOST BOURE[1], Lionel RANJARD[1] and Sébastien TERRAT[1]

[1] Agroécologie – AgroSup Dijon – INRA – Univ. Bourgogne Franche-Comté, 17 Rue Sully, F-21000, Dijon, France

Corresponding Author: sebastien.terrat@inra.fr

Keywords Metabarcoding approaches, OTU definition, Clustering, soil microbial communities

1. Context

Environmental microbial communities are now widely studied using metabarcoding approaches, thanks to the democratization of high-throughput sequencing technologies. The massive number of reads produced with these technologies requires bioinformatic solutions to be efficiently treated. A key step in the analysis is to cluster reads into Operational Taxonomic Units (or OTUs) and thus reduce the amount of data for downstream analyses. Due to the important impact of the method on the quantity and quality of OTUs, finding an equilibrium between the reliability and time-consuming nature of the chosen strategy is a real challenge. Here, we propose a new clustering strategy called ReClustOR which combines two different methods (a *de novo* and a closed- or open- reference method) to overcome some of the problems inherent in many clustering algorithms.

2. Methods

The ReClustOR clustering program was developed in PERL language (v5.26.1 or higher) and comprises two independent modules: the first one is dedicated to definition of the OTUs reference database, and the second one to clustering against a given database.

Firstly, a *de novo* method is used to define OTU centroids and create a reference database. Secondly, a closed- or open-reference method (depending on the user's choice) is computed for all reads which are not considered as OTU centroids. To highlight the improvements provided by ReClustOR in describing microbial diversity in terms of ecological diversity metrics and composition (*e.g.* richness, OTU composition), a massive environmental dataset containing thousands of publicly available samples [1,2] was subjected to a conventional *de novo* method, and to ReClustOR. This dataset focused on the bacterial and archaeal diversity of 1,842 soils samples based on the pyrosequencing of 16S rRNA genes directly amplified from soil DNA. Both methods were analyzed for their ability to efficiently describe microbial richness, and also for the robustness and stability of their OTUs definition.

3. Results

Analysis of the two clustered datasets showed that ReClustOR improves not only the detected richness (Fig 1), but also the reliability and stability of the OTUs, compared to the *de novo* method (Fig 2). More precisely, ReClustOR avoids the accumulation of distant sequences in the same OTU by efficiently comparing all reads to all the centroids of all the defined OTUs. Moreover, ReClustOR, by defining a database of centroids, precludes the need to re-cluster all the reads each time when new reads are generated.

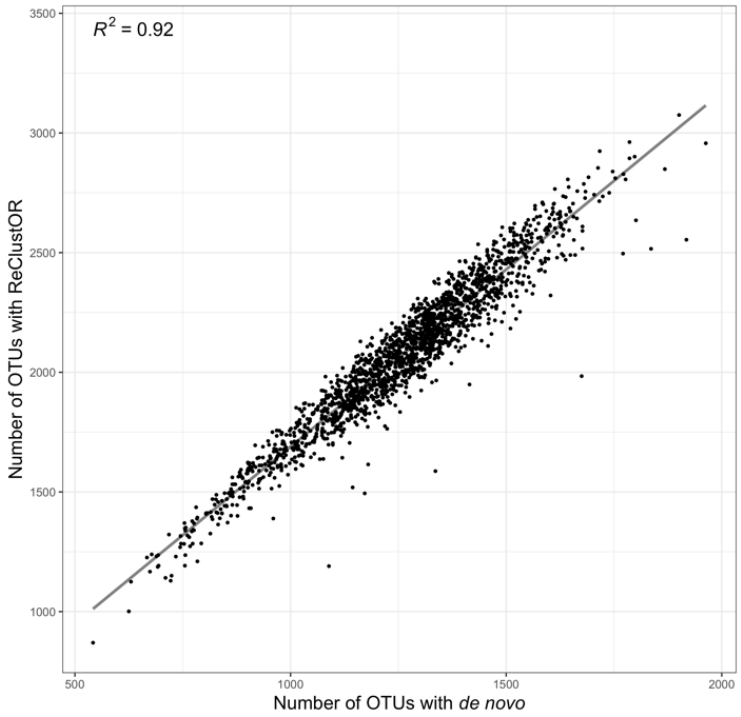


Fig 1. Comparison of the two clustering methods based on microbial richness. The microbial richness observed for each soil sample using the *de novo* versus ReClustOR methods (Pearson's correlation coefficient was computed, p-value < 0.001). The red line represents the unity line.

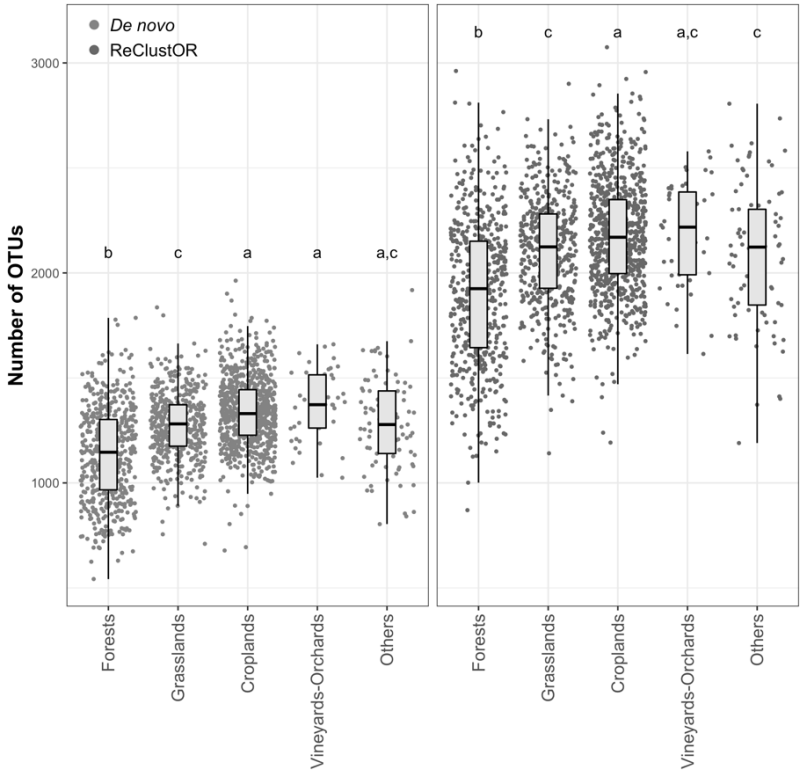


Fig 2. Comparison of observed microbial richness based on land use. Microbial richness for each land use (Croplands, Forests, Grasslands, Vineyards-Orchards and Others) is based on the two clustering methods (*de novo* (left) and ReClustOR (right)) applied to the 1,842 soil samples. Letters indicate significant differences between land uses for each clustering method (ANOVA test, p value < 0.05).

4. Conclusion

ReClustOR is a novel clustering method that overcomes some of the problems associated with classical 'heuristic' clustering methods and consequently increases the stability and quality of the reconstructed OTUs. Moreover, the OTUs database defined with ReClustOR can be used as reference(s) with gradual enrichment of it by merging new studies and samples. In this way, huge datasets like the Earth Microbiome Project can constitute references for other projects within their range of application, thereby increasing the quality of comparisons between studies and datasets, but also improving the extent and the resolution of maps of soil microbial communities.

Acknowledgements

This study was granted by ADEME (French Environment and Energy Management Agency) and by 'France Génomique' through involvement of the technical facilities of Genoscope. Because of the involvement of the technical facilities at the GenoSol platform of the infrastructure Analyses et Expérimentations sur les Écosystèmes (ANAEE) France, it also received a grant from the French state through the National Agency for Research under the program "Investments for the Future" (reference ANR-11-INBS-0001). RMQS soil sampling was supported by a French Scientific Group of Interest on soils: the "GIS Sol," involving the French Ministry of Ecology, Sustainable Development and Energy (MEEM); the French Ministry of Agriculture (MAP); the French Institute for Forest and Geographical Information (IGN); the Environment and Energy Management Agency (ADEME); the French Institute for Research and Development (IRD); and the National Institute for Agronomic Research (INRA).

References

1. Terrat, Sébastien, et al. "Mapping and predictive variations of soil bacterial richness across France." PloS one 12.10 (2017): e0186766.
2. Karimi, Battle, et al. "Biogeography of soil bacteria and archaea across France." Science advances 4.7 (2018): eaat1808.

Recommendation system embedded in metabolic network visualization: a new way of looking at metabolomics results

Clément FRAINAY[1], Maxime CHAZALVIEL[2] and Fabien JOURDAN[1]

[1] INRA UMR1331 Toxalim, 180 Chemin de Tournefeuille, BP 3 31931, Toulouse, France
[2] Medday Pharmaceuticals SA, 24-26 rue de la Pépinière, 75008, Paris, France

Corresponding Author: clement.frainay@inra.fr

Untargeted metabolomics aim at monitoring a large range of metabolites in a sample, which can be used to identify those which concentration is affected by a condition. While it allows to characterize a perturbation, making biological sense out of such data to understand the underlying mechanisms is still a challenging task. It requires to link data to the existing knowledge about the biochemical reactions that embody the relationships between compounds. This knowledge is intuitively represented as a metabolic network[1].

We developed the web server MetExplore[2] (www.metexplore.fr) that gathers various tools to help the task of putting metabolomics results in the context of genome-scale metabolic networks, with a special focus on large network visualization. The visual and interactive exploration of metabolic networks by users, usually aiming to reconstruct metabolic scenarios, is mainly driven by curiosity and usually involves expert knowledge not necessarily modeled in the system, making this task difficult to fully automatize.

On the other hand, such manual exploration is usually tedious without a priori filtering, given the complexity and size of metabolic networks. In order to reduce the information overload, we recently developed a recommendation system inspired by social networks users recommendations, highlighting relevant compounds to investigate based on their connections to markers identified by metabolomics[3].

In order to increase the intelligibility of the recommendations and ease their use in the reconstruction of metabolic scenarios, we embedded them in an interactive network visualization, using our recently published D3.js library dedicated to metabolic networks rendering[4]. We shifted from the traditional "overview then zoom and filter" browsing model for large networks to a bottom-up approach that provides an incremental expansion of users' initial focus[5], empowered by our recommendation system. By highlighting parts of a compound's neighborhood from which known markers can be reached, it assists the reconstruction in a comprehensive way while still providing room to incorporate expert's knowledge at any stage. It thus provides a flexible metabolic network exploration framework managing information overload and prone to serendipitous discoveries.

References

[1] Vincent Lacroix, Ludovic Cottret, Patricia Thébault, and Marie-France Sagot. An introduction to metabolic networks and their structural analysis. In *IEEE/ACM Transactions on Computational Biology and Bioinformatics (TCBB)* 5, no. 4 pages 594-617. 2008.

[2] Ludovic Cottret, Clément Frainay, Maxime Chazalviel, Floréal Cabanettes, Yoann Gloaguen, Etienne Camenen, Benjamin Merlet et al. MetExplore: collaborative edition and exploration of metabolic networks. *Nucleic acids research,* (46):W495–W502, 2018.

[3] Clément Frainay, Sandrine Aros, Maxime Chazalviel, Thomas Garcia, Florence Vinson, Nicolas Weiss, Benoit Colsch et al. MetaboRank: network-based recommendation system to interpret and enrich metabolomics results. *Bioinformatics,* (35):274−283, 2018.

[4] Maxime Chazalviel, Clément Frainay, Nathalie Poupin, Florence Vinson, Benjamin Merlet, Yoann Gloaguen, Ludovic Cottret, and Fabien Jourdan. MetExploreViz: web component for interactive metabolic network visualization. *Bioinformatics,* (34):312–313, 2017.

[5] Frank Van Ham and Adam Perer. “Search, show context, expand on demand”: supporting large graph exploration with degree-of-interest. In *IEEE Transactions on Visualization and Computer Graphics* 15, no. 6 pages 953–960. 2009.

Recurrent deletions of 3q13.31 in human osteosarcoma commonly affect TUSC7 and LINC00901.

Baptiste Ameline*[1], Kovac Michal[1], Karim H. Saba[2], Maxim Barenboim[3], Michaela Nathrath[3,4], Karolin H. Nord[2], and Daniel Baumhoer[1]

[1]Institut of pathology, University Hospital Basel [Basel] – Suisse
[2]Division of Clinical Genetics, Lund University [Lund] – Suède
[3]Pediatric Oncology Center, Technische Universität München, Munich – Allemagne
[4]Pediactric Hematology and Oncology, Klinikum Kassel, Kassel – Allemagne

Résumé

Background: Osteosarcoma, the most common primary tumor of bone, generally harbors complex structural rearrangements. In this study, we identified frequent mono- and bi-allelic deletions of the two long non-coding RNAs TUSC7 and LINC00901, both located on the long arm of chromosome 3 (3q13.31). Their precise biological function remain largely unknown, although it was suggested that TUSC7 may participate in micro-RNA trapping and thereby could be involved in tumor suppressor gene regulation. Moreover, the presence of a TP53 responsive element upstream of their coding region suggests an involvement in the convoluted TP53 signaling pathway.
Methods: Whole genome sequencing data (n=104) and Affymetrix Cytoscan arrays (n=52) were used to derive copy-number profiles of each tumor and interrogated for LINC00901 and TUSC7 deletions. These findings were validated in an independent set of 102 formalin-fixed and paraffin-embedded tissue samples using RNA in situ hybridization techniques.

Results: Mono- and bi-allelic deletions were detected in 68 (43.5%) sequenced and array-profiled osteosarcomas. Of these, 34 tumors (50%) acquired deletions in both genes whilst additional focal deletions of TUSC7 and LINC00901 were acquired in 11 (16%) and 19 (28%), respectively. The four remaining cases (6%) revealed copy number losses within the intergenic region between TUSC7 and LINC00901. In addition, reduced amounts or complete losses of hybridisation signals were detected in similar proportions in the independent set of FFPE samples.
Conclusions: We identified recurrent 3q13.31 deletions in 68/156 (43,5%) human osteosarcomas which seems remarkable in a tumor well known for its high amount of genomic complexity and intertumoral heterogeneity. TUSC7 and LINC00901 might act as downstream effectors of the TP53 pathway and could be functionally equivalent to TP53 inactivation in case of copy number loss. To further elucidate the consequences of 3q13.31 aberrations we aim to correlate our findings with transcriptome data which is currently ongoing.

*Intervenant

Reducing your NGS dataset using a set of targets :
how to optimize storage space, compute time and analysis accuracy

Mathieu Genete
Univ. Lille, CNRS, UMR 8198 - Evo-Eco-Paleo, F-59000 Lille, France
mathieu.genete@univ-lille.fr

The huge amount of data generated by high throughput sequencing technologies (NGS) is a limiting factor for much of computer analysis. Performing these analyzes requires access to infrastructures with sufficient storage space and appropriate computing power. But for some analyzes, only small parts of the initial data are of interest. Reducing the data set, in this case, makes it possible to optimize the storage space, the processor calculation time and the analysis accuracy.
We developed a dedicated algorithm (kmerRefFilter) to reduce, either locally or directly on a download stream, any NGS dataset. This tool produce an exhaustive library of k-mers present in the set of reference sequences, from which duplicates and low-complexity k-mers are removed. This library is next use to filter and reduce the NGS dataset.
We use these NGS sub-dataset for different applications:

- gene assembly of individuals located in a highly polymorphic region, using near or partial targets reference sequences. Such as plant self-incompatibility (SI), a genetic system that prevents selfing and enforces outcrossing. SI are predicted to maintain extraordinary high levels of polymorphism and consequently are typically challenging to assemble de novo as well as to align to a given reference.
- optimisation of variant analysis on targeted genome regions using capture probes as reference to filter and recover matching reads on multiple whole genome sequencing public samples.
- efficient genotyping pipeline to map raw reads from individual outcrossing Arabidopsis genomes against a dataset of multiple reference sequences of the pistil specificity determining gene of the Brassicaceae S-locus (SRK) and determine individual S-genotypes. Drastic reduction of the input dataset, allowing very efficient processing of downstream steps and is adapted for large populations studies. Example of improvement on a 64 cores (Intel Xeon 2.3 Ghz) server:
    - for 5 A. halleri subsp. gemmifera individuals (~ 65 million reads) **without dataset reduction**, compute time is about 5 days
    - for 182 A. halleri subsp. gemmifera individuals (~ 12 billion reads) **with dataset reduction**, only 0,03% of reads of interest are kept and compute time is about 3 hours

For all these applications, despite the reduction of the dataset, all obtained results are just as accurate and show no loss in resolution, with an effective reduction of compute time. This method is very efficient for partial dataset analyse, with a large amount of samples.

Refract-Lyma and CHUN hub: from a research cohort to a regional electronic medical record system and back

Aimeric DABIN[1], Thomas GORONFLOT[1], Benoît TESSOULIN[2], David CHIRON[3], Anne MONLIEN[2], Steven LE GOUILL[2,3], Pierre-Antoine GOURRAUD[1], Matilde KARAKACHOFF[1,4]

[1]CHU de Nantes, INSERM, CIC 1413, Pôle Hospitalo-Universitaire 11: Santé Publique, Clinique des données, Nantes, France

[2]Service d'Hématologie, CHU de Nantes, Nantes, France

[3]CRCINA, INSERM, CNRS, Université d'Angers, Université de Nantes, Nantes, France

[4] l'institut du thorax, INSERM, CNRS, UNIV Nantes, CHU Nantes, Nantes, France

Corresponding author: matilde.karakachoff@univ-nantes.fr

Mantel cell lymphoma (MCL) is a non-Hodgkin lymphoma, a subtype of B-cell lymphoma; a rare and incurable disease that induces patients in continuous relapses and therapeutic resistance. The main challenge is to create innovative tools for the analysis of tumor cell phenotypes and sub-phenotypes, to understand the heterogeneity of the tumor at the time of diagnosis and to determine the mechanisms of response and resistance to treatment. The study of these mechanisms involves the identification of risk and predictive, environmental, genetic and / or behavioral factors that constitute what could be called the "macro-environment" of diseased subjects. The Refract-Lyma (RL) cohort has been set up in order to better characterize this "macro-environment" related to the patient. The CHUN Research Data Warehouse (CHUN hub) has been implemented as an integrated medical and administrative data system hub, containing all the electronic medical records transited by the CHU since 2005 to date and providing a query tool for researchers.

Here we aimed to enrich the RL cohort with real-life data through the implementation of an extraction workflow applied to the CHUN hub. The objective was to produce a set of extraction and identification algorithms for integrating data from the RL cohort on one hand and the data flow contained within the CHUN hub, on the other.

The workflow consisted of two principal blocks. The first block allowed to identify MCL patients from the CHUN hub through the use of extraction algorithm based on text mining rules, applied to both structured and unstructured data, using text keywords, diagnosis and procedure codes, biology exams codes, etc. The second block enabled to match RL cohort patients with the set of individuals extracted in first block; with the purpose of validating the extraction process, analyzing sensitivity and specificity of algorithm rules and assessing reliability and quality of EMR extracted variables.

Preliminary results shown a total of 188 MCL patients extracted using keywords and codes. The sex ratio was 2.3 and median age 75 years, in coherence with figures founded in literature. 44 out of 84 (52%) patients included in RL cohort were identified also in CHUN hub.

This work will contribute with one more piece to the puzzle of clinical “big data” employed in medical research. Advances in the process of enrichment research cohorts with EMR hubs will allow not only to improve data quality and impute missing information but also to help researchers to improve the process of classification and profiling of patients, a fundamental step of predictive medicine.

Acknowledgments

Present work was supported by L'Héma-NexT and Cluster NeXT SysMics, from the i-Site NexT Nantes, France.

REGULOUT software identifies regulatory outliers, that have unexpected transcription profile inside a group of ortholog genes.

Médine BENCHOUAIA[1†], Hugues RIPOCHE[1†], Mariam SISSOKO[1†], Antonin THIÉBAUT[1†], Jawad MERHEJ[1], Thierry DELAVEAU[1], Laure FASSEU[1], Sabrina BENAISSA[1], Geneviève LORIEUX[1], Laurent JOURDREN[2], Stéphane LE CROM[1], Gaëlle LELANDAIS[4], Eduardo COREL[3] and Frédéric DEVAUX[1]

[1] *Sorbonne Université, CNRS, Institut de Biologie Paris-Seine, UMR 7238, Laboratoire de Biologie Computationnelle et Quantitative, Paris, France*

[2] *École Normale Supérieure, PSL Research University, CNRS, Inserm U1024, Institut de Biologie de l'École Normale Supérieure, Plateforme Génomique, Paris, France*

[3] *Sorbonne Université, CNRS, Institut de Biologie Paris-Seine, UMR 7138, Évolution, Paris, France*

[4] *UMR 9198, Institute for Integrative Biology of the Cell, CEA, CNRS, Université Paris-Sud, UPSay, Gif-sur-Yvette, France*

† *These authors have contributed equally to this work*

Corresponding Author: frederic.devaux@sorbonne-universite.fr

In this work, we have developed a new software called REGULOUT.

REGULOUT allows to detect regulatory outliers, i.e. genes which have a particular expression profile inside a group of orthologous genes, from comparative transcriptomics data. The aim is to discover specific regulations of otherwise conserved genes. Given the orthogroups composition and multispecies expression profiles, it calculates, in each orthogroup, the pairwise distances between the expression profiles and use a minimal distance cut-off parameter to identify the genes which expression profile differs from those of all their orthologues and paralogues.

REGULOUT was written in python3 language and its documentation and download area can be found at : http://www.lcqb.upmc.fr/REGULOUT/.

This software has been used for comparative transcriptomics to identify regulatory outliers (ROs)in the human pathogen *Candida glabrata*. ROs are genes that have very different expression patterns compared to their orthologues in other species. From comparative transcriptome analyses of the response of eight yeast species to toxic doses of selenite, a pleiotropic stress inducer, we identified 38 ROs in *Candida glabrata*.

In silico analyses have been confirmed by global chromatin Immunoprecipitation and gene profiling.

References

[1] Médine Benchouaia, Hugues Ripoche, Mariam Sissoko, Antonin Thiébaut, Jawad Merhej, Thierry Delaveau, Laure Fasseu, Sabrina Benaissa, Geneviève Lorieux, Laurent Jourdren, Stéphane Le Crom, Gaëlle Lelandais, Eduardo Corel and Frédéric Devaux, *Comparative Transcriptomics Highlights New Features of the Iron Starvation Response in the Human Pathogen Candida glabrata.* Front. Microbiol. 9:2689. doi: 10.3389/fmicb.2018.02689, 2018.

repeatsFinder: a web-based R/Shiny interface for visualizing and characterize genomic repeated regions

Hugo Varet*[1,2] and Thomas Cokelaer[1,2]

[1]Biomics Pole - C2RT, Institut Pasteur, Paris, France – Institut Pasteur de Paris – France
[2]Hub Bioinformatique et Biostatistique - Bioinformatics and Biostatistics HUB – Institut Pasteur [Paris], Centre National de la Recherche Scientifique : USR3756 – France

Résumé

The presence of repeated regions within a genome may cause problems for de novo assembly. Indeed repeats that are longer than the read length (about 150 bases for short-read technologies) create gaps in the assembly. Similarly, tandem repeats present another common assembly problem since near-identical tandem repeats are often collapsed into a few copies.
New sequencing technologies that generates long reads (e.g., Sequel from Pacific Biosciences or MinION from Nanopore) has become mature and make the genome assemblies more robust thanks to an average read length larger than 10kb (PacBio claims an average read lengths up to 30kb with their newest chemistry as of March 2019). Yet, some genomes have very large repeated regions and others have tandem repeats that are larger than the average read length of long-read technologies. Therefore, it is important to be able to detect long repeated regions that are potentially present in a genome. Indeed, this will allow the experimentalists and bioinformaticians to anticipate the feasibility of an assembly based on the average read lengths of the sequencers and the repetitiveness structure of the genome to be assembled.

We propose a web-based application associated with a comprehensive database that allows to easily visualize the positions and lengths of the repeated sequences of published genomes retrieved from the NCBI website. More precisely, we use the concept of shortest unique substring [1] to define and detect repeated sequences. The website also provides several usual metrics as the proportion of each nucleotide, the genome length, the percentage of the genome covered by repeats or the Ir index [2].

We will present the methodology used to define repetitiveness structure of genomes and the web interface that has been put in place. We will show how this can be used within sequencing platforms for bioinformaticians teams to anticipate the feasibility of robust assemblies.

Haubold B, Pierstorff N, Möller F, Wiehe T. Genome comparison without alignment using shortest unique substrings. BMC Bioinformatics. 2005 May 23;6:123.
Haubold B, Wiehe T. How repetitive are genomes? BMC Bioinformatics. 2006 Dec 22;7:541.

*Intervenant

RGCCA with block-wise missing structure

Caroline PELTIER[1], François-Xavier LEJEUNE[1], Ivan MOSZER[1] and Arthur TENENHAUS[1,2]

[1] Institut du Cerveau et de la Moelle épinière, Inserm U 1127, CNRS UMR 7225, Sorbonne Université, F-75013, Paris, France

[2] Laboratoire des Signaux et Systèmes (L2S), CentraleSupélec, Université Paris-Saclay, F-91190, Gif-sur-Yvette, France

Corresponding Author: caroline.peltier@icm-institute.org

Multidisciplinary approaches are now common in scientific research and provide multiple and heterogeneous sources of measures of a given phenomenon. These sources can be viewed as a collection of interconnected datasets and dedicated algorithms are mandatory for providing relevant information from multi-source data. Regularized Generalized Canonical Correlation Analysis (RGCCA) is a general statistical framework for multi-source data analysis [1]. Multi-source data often have block-wise missing structure, i.e., data in one or more sources may be completely unobserved for a sample. The probability to observe block-wise missing structure increases with the number of sources. It is therefore mandatory to properly handle this block-wise missing structure within the framework of RGCCA.

In this work, several solutions are investigated. A first type of approaches consists in modifying the RGCCA algorithm in order to use only the available data, in the same vein as in Partial Least Squares Path Modeling [2] or using a missing data passive approach [3]. A second type of approaches is based on iterative imputation. Several imputation strategies have been proposed for Principal Component Analysis and Multiple Factor Analysis [4]. We propose to develop iterative imputation within the RGCCA framework. All these methods will be compared on simulations and on multi-source biological data.

Acknowledgements

CP is funded by the iMAP program (ANR-16-RHUS-0001). This work was also partly supported by the IHU-A-ICM program ANR-10-IAIHU-06.

References

[1] Arthur Tenenhaus and Michel Tenenhaus. Regularized Generalized Canonical Correlation Analysis. *Psychometrika*, 76(2): 257–284, 2011.

[2] Michel Tenenhaus, Vicenzo Esposito Vinzi, Yves-Marie Chatelin, Carlo Lauro. PLS path modeling. Computational Statististcs & Data Analysis 48: 159-205, 2005

[3] Michel van de Velden and Yoshio Takane. Generalized Canonical Correlation Analysis with missing values. *Comput Stat* (27): 551-571, 2012

[4] Julie Josse and François Husson. Handling missing values in exploratory multivariate data analysis methods. *Journal de la Société Française de Statistique*, 153 (2): 79-99, 2012

RPG: fast and efficient in silico protein digestion

Nicolas Maillet[1]
Institut Pasteur - Bioinformatics and Biostatistics Hub - C3BI, USR 3756 IP CNRS - Paris, France

Corresponding author: nicolas.maillet@pasteur.fr

Proteases, also known as proteolytic enzymes, have been studied for more than 80 years [1]. Those proteases are widely used in industry, medicine and as a biological research tool, for example in protein characterization or more generally in proteomics and proteogenomics [2]. Recently, interest in proteases has gained importance due to advancements in mass spectrometry techniques used in proteomics and proteogenomics. In "bottom-up" analysis, using tandem mass spectrometry (MS/MS), optimal peptide size range is 600-5,000Da [3] when proteins size are usually more than 10,000 Da. Therefore, for bottom-up approaches, protein digestions are required. To perform digestions, one or several proteases, like trypsin, pepsin or thrombin, are used. Each protease has specific cleavage sites relying on solvent accessibility, pH, temperature, etc. The use of different proteases individually or in combination creates a unique set of peptides. Performing multiple digestions can increase overall confidence in protein identification if cleaving sites are different. It is not always easy to determine which combination of proteases will lead to a set of peptides suitable for MS/MS analysis. However, the cost of some proteases does not allow for easily trying several combinations to avoid redundancy of cleaving sites. Few software exist to predict cleavage sites of proteases in protein sequences. Among those, the most commonly used are PeptideCutter from ExPASy [4] and a module of MaxQuant [5].

This poster presents Rapid Peptides Generator (RPG), a new standalone software dedicated to predict proteases-induced cleavage sites on sequences that overcome some issues existing in commonely used programs. RPG is a python tool taking (multi-)fasta/fastq file of proteins as input and digest each of them. Digestion mode can be either 'concurrent', i.e. all proteases are present at the same time during digestion, or 'sequential'. In sequential mode, each protein will be digested by each protease, one by one. Resulting peptides contain the same informations as PeptideCutter, as-well as an estimation of isoelectric point (pI) of each peptide [6]. Results are outputted in multi-fasta, CSV or TSV file. Currently, 42 proteases and chemicals are included in RPG. User can easily design new proteases, using a simple yet powerful grammar. This grammar allows the user to design complex proteases like trypsin or thrombin, including many exceptions and different cleavage sites. Choosing proteases is not trivial. The same combination of proteases can lead to different results depending on the nature of analyzed proteins. For example, Actin and Globin families reveal different behaviors on two different sets of proteases. One set leads to better results for Actin, the second set for Globin. A third set leads to even better results for both of the families. This highlight that the digestion part of MS/MS analyses should be handle with care and adapted to targeted proteins. RPG can be used to properly define which set of proteases should be used on a specific dataset.

RPG is available through pip ('pip install rpg') and follows the standards for software development with continuous integration on Gitlab (https://gitlab.pasteur.fr/nmaillet/rpg) and automatic on-line documentation (https://rapid-peptide-generator.readthedocs.io).

References

[1] H. Neurath. Proteolytic enzymes, past and future. *Proc. Natl. Acad. Sci. U.S.A.*, 96(20):10962–10963, Sep 1999.

[2] A. I. Nesvizhskii. Proteogenomics: concepts, applications and computational strategies. *Nat. Methods*, 11(11):1114–1125, Nov 2014.

[3] Saveliev S. Urh M. Simpson D. Jones R. Engel, L. and K. Wood. Using Endoproteinases Asp-N and Glu-C to improve protein characterization., 2007.

[4] M. R. Wilkins, E. Gasteiger, A. Bairoch, J. C. Sanchez, K. L. Williams, R. D. Appel, and D. F. Hochstrasser. Protein identification and analysis tools in the ExPASy server. *Methods Mol. Biol.*, 112:531–552, 1999.

[5] J. Cox and M. Mann. MaxQuant enables high peptide identification rates, individualized p.p.b.-range mass accuracies and proteome-wide protein quantification. *Nat. Biotechnol.*, 26(12):1367–1372, Dec 2008.

[6] Tymoczko.J Berg.J and Stryer.L. *Biochemistry 7th edition.* Freeman & Company, W. H., 2011.

RSAT var-tools: an accessible and flexible framework to predict the impact of regulatory variants on transcription factor binding

Walter SANTANA-GARCIA[3], Maria ROCHA-ACEVEDO[1], Yvon MBOUAMBOUA[2], Bruno CONTRERAS-MOREIRA[4], Denis THIEFFRY[3], Morgane THOMAS-CHOLLIER[3], Jacques VAN-HELDEN[2] and Alejandra MEDINA-RIVERA[1]

[1] International Laboratory for Human Genome Research – National Autonomous University of Mexico, Blvd. Juriquilla 3001, 76230, Querétaro, México

[2] Technological Advances for Genomics and Clinics – Aix-Marseille University, 163 Avenue de Luminity, 13288, Marseille, France

[3] Institut de Biologie de l'Ecole Normale Supérieure – Ecole Normale Supérieure, 46 Rue d'Ulm, 75005, Paris, France

[4] European Molecular Biology Laboratory – European Bioinformatics Institute, Genome Campus Hinxton, CB10 1SD, Cambridgeshire, UK

Corresponding Author: amedina@liigh.unam.mx, Jacques.van-Helden@univ-amu.fr

While Genome-wide association studies (GWAS) have successfully pinpointed thousands of genetic variants linked to disease and other traits, there is still not a clear understanding of how most of these variants might be contributing to such complex phenotypes. To date, >71,673 variants have been reported in the NHGRI-EBI GWAS Catalog from which the vast majority are found outside genome coding sequences [1]. Furthermore, regulatory elements reported by ENCODE have shown to be enriched by GWAS-variants, suggesting altered regulatory events as the underlying mechanisms diseases act through.

Transcription Factors (TF) are DNA binding proteins that modulate gene expression by recognizing specific DNA motifs. Genetic variants can modify TF affinity when changes are introduced to bases relevant for DNA-TF contact which could affect the transcriptional regulation. TF binding motifs are usually modeled with Position-Specific Scoring Matrixes (PSSM), which can be used to estimate TF affinity to a specific DNA sequence. Here, we present a series of user-friendly tools to predict the impact of genetic variants on TF binding, accessible through the Regulatory Sequence Analysis Tools (RSAT; http://metazoa.rsat.eu//) [2].

RSAT var-tools is made of four programs: *convert-variations*, *retrieve-variation-seq*, *variation-info* and *variation-scan*. They can be used independently or integrated as a pipeline to address the functional implications of regulatory variants. *convert-variations* interconverts format of variation files between GVF, VCF and varBed, the internal RSAT-specific format for storing variant data. *variation-info* returns annotations about variants given a set of dbSNP IDs or for a given set of genomic coordinates. *retrieve-variation-seq* extracts the variant and flanking sequences from the genome of interest, taking as input either user-supplied variants (varBed) or dbSNP variant ids or dbSNP variants within user-supplied list of coordinates (bed). *variation-scan* assesses the impact of variants on TF binding, estimating and comparing the affinity between a pair of variant alleles. In addition, haplotype phase information can be provided in the VCF file to estimate the joint effect of variants in close proximity on the same TF-binding event. In summary, RSAT var-tools provide a resource to experienced and non expert users (accessible through a web interface) to analyze regulatory variants in several organisms.

References

1. Annalisa Buniello, et al. The NHGRI-EBI GWAS Catalog of published genome-wide association studies, targeted arrays and summary statistics 2019. *Nucleic Acids Research*, 47: D1005–D1012, 2019.
2. Ngathi Thuy Nguyen, et al. RSAT 2018: regulatory sequence analysis tools 20th anniversary. *Nucleic Acids Research*, 46: W209–W214, 2018.

Régulation par les miARN des gènes régulant la fécondité et le développement embryonnaire précoce chez le poisson medaka (*Oryzias latipes*)

Fanny CASSE[1, 2], Emmanuelle BECKER[2], Susete ALVES-CARVALHO[2, 3], Violette THERMES[1], Fabrice LEGEAI[2, 3] et Julien BOBE[1]

[1] LPGP, INRA, Campus de Beaulieu, F-35000 Rennes, France

[2] Univ Rennes, INRIA, CNRS, IRISA, Campus de Beaulieu, F-35000 Rennes, France

[3] IGEPP, INRA BP35327, Le Rheu, France

Corresponding Author: julien.bobe@inra.fr

L'ovogenèse et l'embryogenèse précoce reposent sur des processus biologiques hautement régulés et coordonnés impliquant des interactions géniques et la régulation de gènes. Au cours de l'ovogenèse, les cellules somatiques ovariennes subissent de nombreux changements transcriptionnels afin de préparer les cellules germinales non différenciées à former des gamètes (ovocytes secondaires) [1]. Dans ce contexte, le rôle régulateur des petits ARNs non codant (microARNs) est peu connu. Des études précédentes ont permis de découvrir des microARNs particulièrement exprimés dans l'ovaire, comme miR-202-5p dont le KO entraine une diminution de la quantité et de la qualité des gamètes [2]. Dans le but de mieux comprendre les réseaux moléculaires qui sous-tendent la production de gamètes femelles chez les poissons, nous avons étudié le profil transcriptomique d'ovaires de medaka (*Oryzias latipes*) au cours d'une cinétique d'expression couvrant le cycle de reproduction (T0 : 0h, T1: 4h, T2 : 8h, T3 : 12h, T4 : 16h et T5 : 20h après la ponte journalière). L'alignement des lectures sur le génome à l'aide de STAR [3] , puis leur assemblage par StringTie [4] et leur analyse par FEELnc [5] ont permis de prédire 1131 nouveaux longs ARNs non codants et 539 nouveaux ARN messagers. Ces nouvelles annotations ajoutées à l'annotation de référence *v95 Japanese medaka HdrR* (Ensembl) ont été utilisées pour réaliser les comptages bruts par gène. Puis, 15 contrastes comparant deux à deux les différents points de la cinétique, ont été réalisés avec AskoR [6], un package R simplifiant et automatisant les analyses edgeR, révélant 2412 gènes différentiellement exprimés dans au moins une des comparaisons temporelles, dont 69 nouveaux longs ARNnc et 27 nouveaux ARNm. Un clustering avec la méthode PAM (Partitioning around medoïd, K=11) a permis d'identifier des profils d'expression différentielle pertinents, et suggère une nette différence d'expression entre les premiers temps de l'ovogénèse (T0 à T3) et les temps plus tardifs (T4 et T5). Notre étude se poursuit à l'heure actuelle par une caractérisation fonctionnelle des clusters (notamment par un enrichissement en termes de la Gene Ontology, ou en cibles de miARNs s'exprimant principalement dans l'ovaire).

References

[1] T. Iwamatsu, « Stages of normal development in the medaka Oryzias latipesq », *Mechanisms of Development*, p. 14, 2004.

[2] S. Gay *et al.*, « MiR-202 controls female fecundity by regulating medaka oogenesis », *PLOS Genetics*, vol. 14, n° 9, p. e1007593, sept. 2018.

[3] A. Dobin *et al.*, « STAR: ultrafast universal RNA-seq aligner », p. 7.

[4] M. Pertea, G. M. Pertea, C. M. Antonescu, T.-C. Chang, J. T. Mendell, et S. L. Salzberg, « StringTie enables improved reconstruction of a transcriptome from RNA-seq reads », *Nature Biotechnology*, vol. 33, n° 3, p. 290-295, mars 2015.

[5] V. Wucher *et al.*, « FEELnc: a tool for long non-coding RNA annotation and its application to the dog transcriptome », *Nucleic Acids Research*, vol. 45, n° 8, p. 12, 2017.

[6] https://github.com/askomics/askoR

Réseaux de co-expression pour l'analyse de données de protéomique pour la compréhension des mécanismes d'action de contaminants chez une espèce non-modèle, *Gammarus fossarum*.

Natacha Koenig[1], Christine Almunia[2], Arnaud Chaumot[1], Jean Armengaud[2], Olivier Geffard[1], Davide Degli Esposti[1]

1 Irstea, UR RIVERLY, Ecotoxicology Group, 5 rue de la Doua, CS 20244, F-69625, Villeurbanne Cedex, France

2 CEA-Marcoule, DRF/Joliot/DTMS/SPI/Li2D, Laboratory "Innovative technologies for Detection and Diagnostics", BP 17171, F-30200, Bagnols-sur-Cèze, France.

Corresponding Author: davide.degli-esposti@irstea.fr

Grâce aux nouvelles technologies de séquençage et de spectrométrie de masse, l'accès aux transcriptomes et protéomes est rendu possible pour des espèces non-modèles dont le génome n'est pas disponible. L'utilisation des approches –omiques s'est élargie à de nouveaux domaines tels que l'écotoxicologie avec l'objectif d'informer sur les mécanismes d'actions des contaminants chez des espèces de pertinence environnementale, sur les déterminismes moléculaires de la sensibilité des populations naturelles ou encore de pouvoir identifier des biomarqueurs d'effet et d'exposition utilisable par exemple en biosurveillance. Les approches de réseaux de co-expression peuvent aider à extraire et sélectionner les informations biologiques pertinentes issues des jeux de données –omiques, notamment dans le contexte d'organismes non-modèles pour lesquels l'annotation des protéines n'est pas encore aboutie. En écotoxicologie, ces mêmes approches peuvent permettre d'étudier les voies moléculaires affectées par l'exposition aux contaminants, et de les corréler aux effets toxiques. Dans le cadre de notre étude, nos objectifs ont été i) d'adapter une méthode de construction de réseaux de co-expression classiquement utilisées pour l'interprétation de données de puces à ADN (package R Weigthed Gene Co-expression Network Analysis : WGCNA) à des données de protéomique shotgun, ii) d'identifier les modules de protéines co-exprimées dans un contexte d'exposition au laboratoire de l'espèce de crustacé sentinelle *Gammarus fossarum* à des substances chimiques modèles (un métal et deux insecticides) connues pour leur reprotoxicité potentielle chez les arthropodes et induisant une réduction de la production de spermatozoïdes chez le gammare, iii) d'établir des corrélations entre les modules et l'exposition aux contaminants testés.

L'analyse a été effectuée sur un jeu de données de protéomique shotgun, constitué de 40 échantillons préparés à partir des testicules de gammares exposés à deux concentrations de cadmium (Cd), pyriproxyfène (Pyr) et méthoxyfénoside (Met). La comparaison de différentes méthodes de normalisation a permis d'identifier la méthode de normalisation la plus adaptée aux données de protéomique. Une analyse de réseaux de co-expression basée sur le package R WGCNA a été ensuite réalisée.

La comparaison des méthodes de normalisation a montré que la méthode la plus adaptée est la procédure Trimmed Mean of M-values (TMM) provenant du package EdgeR conçue pour des données de comptages issues de données RNA-seq. L'analyse de réseau a mis en évidence six modules de protéines co-exprimées. Parmi ces derniers, trois modules distincts ont été identifiés comme significativement corrélés à chaque contaminant. L'analyse d'enrichissement a permis d'identifier pour chaque module associé aux trois substances des protéines impliquées dans des processus biologiques spécifiques, suggérant ainsi différents mécanismes d'action sous-jacents à l'infertilité induite par l'exposition à ces trois substances.

Cette analyse a montré que les réseaux de co-expression sont des outils performants et adaptés pour exploiter les données issues de la protéomique shotgun chez *Gammarus fossarum*, et ce même en l'absence d'un génome annoté. Ces approches aident à mettre en lumière les mécanismes d'actions des contaminants et identifier des acteurs moléculaires nécessitant des analyses fonctionnelles approfondies.

Sex-specific differences in microglia inflammatory response during brain development

Authors:

Lakoum S (1), Touibi N (1), Van Steenwinckel J (1), Rangon CM (1), Fleiss B (1,2), Gressens P (1,2), Delahaye-Duriez A (1,3,4)

1. NeuroDiderot, UMR1141, INSERM, Université Paris Diderot, Sorbonne Paris Cité, Paris, France

2. Centre for the Developing Brain, Department of Perinatal Imaging and Health, Division of Imaging Sciences and Biomedical Engineering, King's College London, King's Health Partners, St. Thomas' Hospital, London, UK

3. UFR SMBH, Université Paris 13, Sorbonne Paris Cité, Bobigny, France

Many epidemiological studies have shown sex-specific differences for many neurodevelopmental and neurodegenerative diseases such as autism spectrum disorder (ASD), schizophrenia, Alzheimer's disease, Parkinson's disease and other several autoimmune diseases [1]. For most of these pathologies, neuroinflammation is a common denominator. Microglia, as the macrophage of the central nervous system, is one of the main cells in neuroinflammatory process. In humans as in mice, the distribution and functions of this type of cells within the central nervous system are regulated during the brain development.

In many papers published recently, authors studied the fact that the sex differences observed for many neurodevelopmental diseases are at least partially related to microglia [2].

In this project we tried to find out if the transcriptional response to inflammatory stress induced at an early stage of development differs according to the sex. The purpose is to characterize the biological pathways and the transcriptional signatures of microglia that are specific to each sex.

Here, we used a validated murine model where a perinatal inflammation was induced by systemic interleukin-1β (IL-1B) administration in two groups (males/females) at the postnatal days P1 and P2, and last injection at P3 just two hours before isolating microglia cells from the mouse brains [3].

RNAseq data were generated from these isolated cells with 6 biological replicates for each sex. Next we performed a differential gene expression analysis using DESeq2 package [4] first to compare systemic IL-1B exposure

with control conditions in each sex (male and female), and then to examine transcriptional changes in both sexes, in inflammation and control conditions. A functional analysis was also performed using Gene set enrichment analysis (GSEA) method [5] which is based on a ranked list of gene scores that take into account not only the differential expression but also the significance of this expression.

At P3, expression changes in microglia after systemic exposure to IL-1B in females were significantly correlated to those observed in males (Spearman correlation coefficient 0.83) .The pathways significantly enriched in expression changes in microglia after *in vivo* systemic exposure to IL-1B are concordant with those obtained from differential expression and GSEA in mouse primary microglial cell cultures under *in vitro* inflammation stimulation by exposure to IL-1B and interferon-gamma. Comparison with the expression changes after systemic IL-1B exposure in males of the same mouse model but at other times of development (P1, P5, P10 and P45 [6]) showed the same pathways are similarly down or up regulated in the different developmental time points in males.

On the other hand, focusing on the expression changes between both sexes, the GSEA showed opposite directions in IL-1B stimulation of inflammation compared to physiological condition for several GO biological pathways (GO Extracellular matrix).

To conclude, our results suggest that the microglia transcriptional response to inflammation is strongly regulated mainly similarly in males and females.

Next steps of this project include gene co-expression-based analyses with topological measures to pinpoint gene master regulators of sex-specific inflammation-associated modules. A better understanding of the sex-specific molecular mechanisms that regulate the microglia activation in the developing brain may help to develop effective therapies to prevent brain damage and neurodevelopmental disorders.

References :

[1] Salter et al Nature Medicine (2017) [2]Hanamsagar et al. Glia (2017) [3] Favrais et al. Ann neurol(2011) [4] Michael I Love,Wolfgang Huber and Simon Anders, Genome Biology(2014) [5] Subramanian et al. PNAS (2005) [6] Krishnan et al. Nature Communication(2017)

Simulating the impact of Serological-Test-and-Treat measures to target the hidden *P. vivax* reservoir: public health impact and primaquine overtreatment

Thomas OBADIA[1,2], Michael T. WHITE[1], Narimane NEKKAB[1] and Ivo MUELLER[2]

[1] Malaria: Parasites and Hosts, Department of Parasites and Insect Vectors, Institut Pasteur, Paris

[2] Hub de Bioinformatique et Biostatistique – C3BI, Institut Pasteur, USR 3756 CNRS, Paris

[3] Population Health & Immunity Division, Walter and Eliza Hall Institute of Medical Research, Parkville, Australia

[4] Department of Medical Biology, University of Melbourne, Melbourne, Australia

Corresponding Author: thomas.obadia@pasteur.fr

Plasmodium vivax remains a major cause of malaria outside of Africa. Reaching pre-elimination conditions remains challenging because of the undetectable reservoir of liver-stage hypnozoites responsible for relapses. Mathematical models can account for this particular biology and allow prediction of the impact of control measures.[1] Recent advances in serological markers of exposure are driving the development of diagnostic tests that can identify likely hypnozoite carriers by measuring the antibody response to recent infections.[2] This enables Serological-Test-And-Treat (STAT) approaches to only target likely hypnozoite carriers for primaquine treatment, compared to a Mass Drug Administration (MDA) where the entire population is exposed to primaquine with potential for toxicity.

Using the models from White *et al.,*[1] we simulated the impact of STAT with 14 day primaquine and testing for G6PD deficiency in a population resembling that found in Papua New-Guinea (PNG), subject to varying *P. vivax* transmission (qPCR prevalence resp. <1%, ~5% and ~10%). While current antibody panels allow for detection of likely hypnozoite carriers with 80% sensitivity and 80% specificity, we explored a wider parameter space to assess the potential public health benefits compared to risk of overtreatment with hypnozoiticidal drugs such as primaquine. The impact of control strategies is measured as the reduction in qPCR *P. vivax* prevalence 6 months after intervention; primaquine overtreatment is defined as administration of primaquine that was either inefficient or unnecessary.

An MDA program at 80% coverage is predicted to cause a 50% - 70% reduction in *P. vivax* PCR prevalence. A single round of SSAT with 80% sensitivity and 80% specificity would lead to a 45% - 75% reduction in prevalence, performing nearly as well as MDA with the benefit of reducing the number of people overtreated with primaquine by 80%. In all prevalence scenarios, increasing sensitivity increased the public health impact while primaquine overtreatment remained constant. Conversely, increased specificity resulted in fewer people receiving unnecessary primaquine, with minor reductions in public health impact.

Population-based treatment strategies with primaquine need to balance the public health impact versus the risk of primaquine overtreatment (treating people that do not have hypnozoites, or to whom primaquine will either be inefficient or induce toxicity). This is mirrored in the balance between sensitivity and specificity of serological diagnostic tests. Improving both sensitivity and specificity will allow us to obtain the best of both worlds: targeting primaquine at people who need it, without endangering those who don't need it.

References

1. White, M. T. *et al.* Mathematical modelling of the impact of expanding levels of malaria control interventions on Plasmodium vivax. *Nat. Commun.* **9**, 3300 (2018).
2. Longley, R. J. *et al.* Development and validation of serological markers for detecting recent exposure to Plasmodium vivax infection. *bioRxiv* 481168 (2018). doi:10.1101/481168

Single cell transcriptomic analysis for a better understanding of human CD8 regulatory T cells

Céline Sérazin[1, 2], Léa Flippe[1, 2], Dimitri Meistermann[1, 2], Séverine Bézie[1, 2], Laurent David [1, 2], Carole Guillonneau[1, 2]

[1]INSERM UMR1064, Center for Research in Transplantation and Immunology ITUN, Université de Nantes, 30 Bd. Jean Monnet, 44093 Nantes Cedex 01, France

[2]Institut de Transplantation Urologie Néphrologie (ITUN), CHU Nantes, 44000 Nantes, France

Corresponding Author: carole.guillonneau@univ-nantes.fr

Organ or cell transplantation is the only therapeutic solution for pathologies causing an irreversible loss of vital organs function. One major goal in transplantation is to develop novel specific and non-toxic anti-rejection immunotherapies. Strategies based on regulatory T cells (Tregs) are promising. Tregs are known to be able to prevent graft rejection. We focus on $CD8^+CD45RC^{low}$ Tregs that have shown suppressive function in vitro and in vivo in rat and human. We have shown that cell therapy using human $CD8^+CD45RC^{low}$ Tregs was efficient to prevent graft rejection and GVHD in immunodeficient NSG mice [1]. However, the heterogeneity of the $CD8^+CD45RC^{low}$ Tregs population is important from a phenotypic point of view, suggesting that either a fraction of the population is tolerogenic, or the induction of tolerance is due to a combination of cells forming an "immunological niche". In order to be able to discriminate between these two hypotheses, we explored the heterogeneity of the $CD8^+CD45RC^{low}$ cell population.

We isolated human $CD8^+CD45RC^{low}$ cells from several healthy volunteers and studied them by single cell RNA sequencing methods using two technologies: Fluidigm C1 and 10X Genomics and compared to public dataset of human PBMCs. The analysis with Seurat package in R highlighted the heterogeneity inside the population with the formation of 4 distinct clusters explaining from 14% to 35% of the whole population. One cluster only was associated with the expression of classical tolerogenic molecules associated with a regulatory function and several new markers. Ten of these news markers were further assessed for their expression by flow cytometry and suppressive test in vitro.

This project will provide crucial information on the biology of $CD8^+$ Tregs in humans and new perspectives in human transplantation.

Key words: immunotherapies, single cell RNA sequencing, CD8 regulatory T cells.

References

[1] - Bézie S, Meistermann D, Boucault L, Kilens S, Zoppi J, Autrusseau E, Donnart A, Nerrière-Daguin V, Bellier-Waast F, Charpentier E, Duteille F, David L, Anegon I and Guillonneau C (2018*). Ex Vivo Expanded Human Non-Cytotoxic CD8+CD45RClow/− Tregs Efficiently Delay Skin Graft Rejection and GVHD in Humanized Mice. Front. Immunol. 8:2014. doi: 10.3389/fimmu.2017.02014*

Single-cell analysis of human intestinal organoids reveals the ENS progenitor cells contribution on the gut mesoderm development.

Authors: Elise Loffet[1], Nicole Brown[2], Nambirajan Sundaram[2], Carine Bouffi[2], Holly M Poling[2], Michel Neunlist[1], Michael A Helmrath[2], Maxime M. Mahe[1,2]

[1]Inserm UMR 1235 - TENS, University of Nantes, Inserm, 1 Rue Gaston Veil, 44035 Nantes Cedex 1, France
[2]Division of Pediatric General and Thoracic Surgery, Cincinnati Children's Hospital Medical Center, Cincinnati, OH, USA

Corresponding Author: elise.loffet@univ-nantes.fr

Introduction: The enteric nervous system (ENS) is one of the key regulators of the intestinal cellular microenvironment. However, its role in the maturation of the developing digestive tract remains poorly understood. ENS progenitor cells (vagal neural crest cells or vNCCs) and gut endoderm co-develop during gut formation. In this context, we hypothesized that the ENS contributes to gut development by influencing patterning of mesodermal cells.

Methods: HIOs are human intestinal tissue produced *in vitro* from directed differentiation of human pluripotent stem cells (hPSCs). The HIO model can also include hPSC-derived enteric neuroglial cells (HIO+ENS) to obtain innervated intestinal organoids. To address our hypothesis, we used human intestinal organoids (HIO) with and without an ENS. After generating HIO and HIO+ENS, we performed differential gene expression (R package DESeq2) and gene ontology analysis of differentially expressed genes using bulk RNA sequencing. We then performed a canonical correlation analysis and alignment of HIO and HIO+ENS single-cell RNA sequencing data (R package Seurat) to identify cell types that are conserved or different across our conditions.

Results: Differential gene expression and ontology analysis on bulk RNA sequencing of intestinal organoids demonstrated that HIO+ENS present increased expression of mesoderm and derivative tissues. In addition, single cell sequencing results demonstrated novel mesodermal populations in HIO+ENS.

Conclusion: These results strongly suggest that ENS progenitor cells impact the development and patterning of intestinal mesoderm-derived cells. This needs to be confirmed by further experimentation to decipher the cell fate transitions and the mechanisms involved in the gut mesoderm patterning using advance single cell modalities.

JOBIM 2019 Poster

SIStemA : Gene expression database of human Stem Cell and their differentiated derivative.

Margot JARRIGE[1], Hélène POLVÈCHE[1], DIDIER AUBOEUF[2], CÉCILE MARTINAT[3], MARC PESCHANSKI[1,3]

[1] CECS / I-Stem, AFM, Corbeil-Essonnes, 91100, France
[2] LBMC, 46 Allée d'Italie, 69364 Lyon Cedex 07, France
[3] INSERM, UMR 861, UEVE, ISTEM, AFM, 91100 Corbeil-Essonnes, France.

Corresponding Author: hpolveche@istem.fr

I-Stem is a French laboratory for research and development on rare monogenic disease using human pluripotent stem cells (hESC/hIPSC). These cells, which are able to self-renewal and to differentiate into any cell type, has emerged as a powerful tool for disease modelling, drug screening and cell therapies. Next Generation Sequencing allows to understand the way of genes are transcribed and regulated on a particular cellular context.

Since 5 years, our platform has sequenced more than 350 dataset of hESC/IPSC derived cell with an amplicon-based enrichment method covering the expression of 21080 genes (Ampliseq ™). This samples were analysed with a specific pipeline according to this technology and the genes expression was stored in a MySQL database (will be available for download).

An user-friendly web interface was developped, called SIStemA and allows to choose cell criteria and genes set to visualize easily their co-expression. The aim of this tool is to answer questions such as how much genes of interest are transcribed in a specific hESC/hIPSC-derived cell type and how their expression evolves in a diseased state or in a given experimental condition.

Keyword : *human pluripotent stem cells, rare monogenic disease, Ampliseq, database, interface*

SpecOMS: découverte des modifications portées par les protéines

Dominique Tessier[1], Matthieu David[1,2], Virginie Lollier[1], Guillaume Fertin[2] et Hélène Rogniaux[1]

[1] INRA UR1268 Biopolymères Interactions Assemblages, F-44316 Nantes, France

[2] LS2N UMR CNRS 6004, Université de Nantes, F-44300, Nantes, France

Corresponding Author: Dominique.Tessier@inra.fr

1. Contexte

L'univers des protéines reste encore très largement méconnu : 99.8% des protéines décrites dans la banque de référence Uniprot sont prédites *in silico* à partir de l'information génomique disponible et donc non strictement identifiées. La spectrométrie de masse en mode MS/MS est la technique majoritairement utilisée pour caractériser les protéines. La qualité des résultats d'une analyse dépend bien sûr de ses conditions expérimentales mais aussi, pour une large part, de la capacité des logiciels à interpréter les dizaines de milliers de spectres générés. Cette étape d'interprétation est chronophage et le taux d'interprétation des spectres n'est encore que d'environ 25%. Ce faible taux est communément expliqué par la présence de modifications portées par les protéines et non connues a priori. Ces modifications peuvent correspondre à des modifications post-traductionnelles essentielles pour l'activité des protéines, à des variants pouvant expliquer certains phénotypes ou pathologies, ou encore à des artefacts liés à la préparation des échantillons.

En renouvelant le paradigme de la comparaison de grands volumes de spectres, nous avons développé le logiciel SpecOMS [1,2], qui permet de comparer des dizaines de milliers de spectres expérimentaux à des centaines de milliers de spectres modélisés à partir d'une banque de protéines, ceci en quelques minutes sur un poste de travail standard. Cette rapidité permet ainsi de s'affranchir du filtre de masse habituellement utilisé pour limiter le nombre de comparaison entre les spectres, filtre qui exclut la recherche de la plupart des modifications.

2. Résultats

Nous illustrons l'intérêt de SpecOMS sur un jeu de spectres téléchargé depuis la banque PRIDE [3] (PXD004732 [4]) au travers de trois contextes d'utilisation différents. Tout d'abord, SpecOMS est parfaitement adapté à un contrôle très rapide des artefacts d'une expérimentation avec une représentation graphique dédiée. Ensuite, sans filtre de masse, SpecOMS met en évidence à une large variété de modifications portées par les protéines en une seule analyse, des plus fréquentes aux plus rares. Enfin, nous montrons que grâce aux performances des nouveaux spectromètres de masse, l'usage d'un très grand espace de recherche pour identifier chaque spectre ne dégrade pas la sensibilité par rapport aux logiciels conventionnels : la découverte de nouvelles modifications ne se fait pas au dépend d'une perte du nombre d'identifications.

References

1. David, M., Fertin, G., Tessier, D. & . SpecTrees: an efficient without a priori data structure for MS/MS spectra identification. in *16th Workshop on Algorithms in Bioinformatics (WABI 2016)* 65-76 (Springer-Verlag, Aarhus, Denmark, 2016).
2. David, M., Fertin, G., Rogniaux, H. & Tessier, D. SpecOMS: A Full Open Modification Search Method Performing All-to-All Spectra Comparisons within Minutes. *J Proteome Res* **16**, 3030-3038 (2017).
3. Jones, P., *et al.* PRIDE: a public repository of protein and peptide identifications for the proteomics community. *Nucleic Acids Res* **34**, D659-663 (2006).
4. Zolg, D.P., *et al.* Building ProteomeTools based on a complete synthetic human proteome. *Nat Methods* **14**, 259-262 (2017).

srnaMapper: a mapping tool for short RNA reads

Matthias Zytnicki and Christine Gaspin
Unité de Mathématiques et Informatique Appliquées, Toulouse INRA, Auzeville BP 52627 31326 CASTANET TOLOSAN cedex FRANCE

Corresponding author: matthias.zytnicki@inra.fr

1 Introduction

Very short RNAs (sRNAs) are key element of genomic regulation [1]. Their sizes usually range from 20 to 30 nucleotides, and belong to the larger class of non-conding RNAs. Recently, most of the research in the field focused on microRNAs (miRNAs), as they have been shown to be implicated in cell development and differentiation, among others. However, other sRNAs, such as piwi-interacting RNA (piRNAs), small interfering RNA (siRNAs), and tRNA-derived RNA fragments (tRFs), to name a few, have also been characterized, with many diverse roles, espicially in epigenetic regulation.

sRNA sequencing (sRNA-Seq) produces evidence of the presence of sRNAs in a given genomic context. Although the protocol is quite similar to RNA sequencing, the output is radically different. First, the size of the transcripts is usually very short. Second, since some miRNAs and tRFs are highly expressed, the same sequence may be sequenced numerous times, providing highly redundant datasets. Third, some sRNA families repress transposable elements, and are thus likely to be present at numerous *loci* in the genome. Last, some miRNAs are edited, and their ends may be replaced, appended, or removed.

Mapping, *i.e.* predicting the possible *loci* that produced a read, is one of the first step of the sRNA-Seq analysis pipeline. To date, there is no dedicated tool for sRNA-Seq, and very few algorithmic effort has been put to these very short reads. Users usually rely to DNA-Seq mapping tools, such as `bwa` [2]. However, these tool do not handle the specificities of sRNA-Seq.

2 Results

We implemented a prototype of a new mapping tool. The strategy is the following.

First, we store the reads into a suffix tree. Identical reads are merged, but the counts are kept, and the quality is defined as the maximum base-wise quality.

Next, the tree is compared to the suffix array (using the Burrows–Wheeler transform and the FM index) of the genome. The aim is to find, for each cell of the suffix tree, the set of the suffix array intervals with minimum edit distance to the cell. Our algorithm recursively explores the tree using a depth-first traversal, and uses the information of the parent cell to find the corresponding suffix array intervals, akin to dynamic programming. To accelerate the search, the algorithm first tries to map with no error. If this search fails, it adds an error, find the corresponding suffix array intervals, etc.

We used the `bwa` API to implement the suffix array operations. As a result, the genome databases built for `bwa` can be used by our method.

In the current implementation, we can align slightly more reads than `bwa mem`, at the expense of speed. However, we expect to improve our algorithm soon.

The code is available on: https://github.com/mzytnicki/srnaMapper.

References

[1] Axtell M. Classification and Comparison of small RNAs from Plants. *Annual Review of Plant Biology*, 64:137–159, 2013.

[2] Li H and Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics*, 25:1754–1760, 2009.

Statistical inference of immunogenetic parameters reveals an *HLA* allele associated with pediatric Focal Segmental Glomerulosclerosis

Axelle DURAND[1,2], Cheryl WINKLER[3], Nicolas VINCE[1,2], Venceslas DOUILLARD[1,2], Estelle GEFFARD[1,2], Derek K. NG[4], Pierre-Antoine GOURRAUD[1,2], Bradley WARADY[5], Susan FURTH[6], Jeffrey B. KOPP[7], Frederick J. KASKEL[8] and Sophie LIMOU[1,2,9]

[1] CRTI UMR 1064, INSERM, Université de Nantes, Nantes, France

[2] Institut de Transplantation Urologie Néphrologie (ITUN), CHU Nantes, Nantes, France

[3] Frederick National Laboratory, Leidos Biomedical Research, NIH/NCI, Frederick MD, USA

[4] Johns Hopkins Bloomberg School of Public Health, Baltimore MD, USA

[5] Children's Mercy, Kansas City MO, USA

[6] Children's Hospital of Pennsylvania, Philadelphia PA, USA

[7] NIDDK, NIH, Bethesda MD, USA

[8] Einstein/Montefiore, Bronx NY, USA

[9] Ecole Centrale de Nantes, Nantes, France

Corresponding Author: Sophie.limou@univ-nantes.fr

Abstract *Focal segmental glomerulosclerosis (FSGS) is a major cause of pediatric nephrotic syndrome, which is often characterized by edema and massive proteinuria. In many cases, it progresses to kidney failure for which treatment is renal replacement therapy by dialysis or kidney transplant. African Americans exhibit an increased risk for developing FSGS, and predisposing genetic factors have been described. Here, we conducted the first genomic study of renal failure in children of African ancestry in order to identify its underlying genetic determinants. DNA from 140 African American children with chronic kidney disease, selected from the CKiD cohort was genotyped on Illumina Exome chips covering >254,000 SNPs. After quality control and SNP imputation, 934,000 common SNPs (minor allele frequency≥0.3%) were tested for association with renal phenotypes using regression models. SNP association analysis revealed 67 SNPs from 5 genes significantly associated with FSGS (FDR<5%). Among these, we highlighted genetic variants within the APOL1 gene ($P=8.6x10^{-7}$, OR=25.4) and the ALMS1-NAT8 locus ($P=1.2x10^{-7}$, 3/29 cases vs. 0/125 controls), which had been previously associated with adult FSGS and chronic kidney disease. Interestingly, we also identified associations with FSGS for the PTPRJ gene ($P=3.4x10^{-7}$, 3/27 cases vs 0/97 controls), that encodes for a signaling receptor interacting with class II HLA molecules. A gene-set enrichment analysis confirmed the importance of antigen processing and presentation pathways in pediatric FSGS ($P=1.3x10^{-6}$). To further the investigation of HLA, we imputed 108 HLA alleles from SNPs using the machine-learning HIBAG tool and inferred additional immunogenetic parameters (HLA amino acids and haplotypes) with Easy-HLA. The strongest associations with FSGS were found for HLA-DRB1*11:01 ($P=5.6x10^{-3}$, OR=10.5) and the 67F and 58E HLA-DRB1 amino-acids ($P=5x10^{-3}$, OR=4.5). To conclude, this first genomic investigation of pediatric FSGS in African American children identified five biologically-relevant, statistically significant loci and highlighted a role for class II HLA molecules in the molecular pathogenesis. Further genetic and functional analyses focusing on these loci will enhance our understanding of molecular mechanisms underlying pediatric FSGS.*

Keywords Chronic Kidney disease (CKD), SNP, FSGS, GWAS, HLA, GSEA, bioinformatic exploration.

Stratégie de compression de données de séquençage cliniques

Laureline Dejardin Bretones, Aubin Thomas and William Ritchie
CNRS - IGH, 141 Rue de la Cardonille, 34090 Montpellier, France

Auteur référent: laureline.dejardin-bretones@etu.umontpellier.fr

Abstract *Les grandes avancées technologiques ont fait basculer la recherche médicale dans l'ère du Big Data. Les méthodes actuelles de modélisation, de traitement et d'analyse ne sont dorénavant plus adaptées face au déluge de données. De ce fait de nouvelles méthodes spécialisées émergent dans le traitement de la donnée médicale. De façon commune à ces nouvelles approches la structuration et l'indexation des données sont des enjeux primordiaux et permettent une faisabilité de l'analyse de cohortes de patients. La modélisation en k-mers des données de séquençage de patients se révèle moins biaisée mais est encore plus coûteuse en temps de calcul. Ce manuscrit propose une étude de différentes méthodes de structuration, d'indexation et de compression rendant exploitable l'analyse de données en k-mers pour de larges cohortes de patients.*

Keywords Indexation, compression, requêtage, aglorithme.

Stratégie de priorisation de variants après séquençage ciblé de l'ADN

Emmanuel Gilson[1], Ségolène Diry[1], Pierre Sujobert[2], Kaddour Chabane[2], Alban Ott[1], Eric Ginoux[1], and Virginie Chesnais*[1]

[1]LifeSoft – Entreprise privée – France
[2]Centre de Recherche en Cancérologie de Lyon – Centre National de la Recherche Scientifique : U1052 – France

Résumé

Keywords NGS, variants prioritization, machine learning

Introduction

La démocratisation du séquençage à haut-débit a comme conséquence l'accumulation d'une quantité importante de données nécessitant une analyse approfondie. La bioinformatique permet de répondre à ce besoin depuis le traitement des données brutes en sortie de séquençage jusqu'à l'interprétation des résultats. Dans le domaine de l'oncologie notamment, le séquençage des tumeurs permet de générer des profils mutationnels qui participent aux décisions quant à la prise en charge des patients. Ainsi, il est nécessaire de générer des pipelines permettant de i. filtrer et aligner les séquences de bonne qualité, ii. Réaliser les appels de variants, iii. annoter ces variants selon différents critères et iv. hiérarchiser les variants selon ces annotations. Cette dernière étape est cruciale : c'est elle qui va permettre de définir quels sont les variants d'intérêt. Il s'agit le plus souvent d'appliquer des critères plus ou moins stricts afin de réduire la liste des variants d'intérêt. Cependant cette approche nécessite de cacher des variants à l'utilisateur final ce qui peut être problématique et est difficilement portable car les caractéristiques d'un variant d'intérêt peuvent être différentes selon les projets. Face à ces limites, nous avons choisi une approche basée sur un algorithme de machine learning, exploitant l'information contenue dans les annotations générées par le pipeline bioinformatique pour construire une frontière de décision séparant les variants d'intérêts du reste.

Résultats

Dans un premier temps, nous avons construit un pipeline qui permet le traitement et l'annotation des données de séquençage. Pour l'appel de variants, nous avons évalué 6 outils différents afin de déterminer les 3 plus performants. Une série d'annotation est alors réalisée afin de caractériser précisément les variants identifiés. Au final nous totalisons 25 annotations de type continue ou catégorielle et pouvant être classées en 3 groupes principaux : les *annotations qualitatives* telles que la qualité des bases alternatives, la profondeur à la position et des métriques concernant l'environnement génomique des variants ; les *annotations fonctionnelles* telles que la présence des variants dans des bases de données publiques et des scores de prédiction de l'impact sur la protéine ; les *annotations quantitatives* telle que les fréquences alléliques des variants.

*Intervenant

Notre jeu de données a été obtenu après séquençage de cellules de moëlle osseuse de patients porteurs d'une leucémie aigüe. Nous avons analysé les résultats de 61 échantillons et ainsi obtenu une moyenne de 1222 variants candidats par échantillon. Un premier filtre permettant d'éliminer les variants introniques et synonymes nous a permis de réduire la liste à une moyenne de 518 variants candidats par échantillon. Parmi eux, 0 à 13 variants par patients ont été annotés comme positifs par une analyse manuelle et nous ont servis de jeu d'entrainement et de validation pour la mise en place de notre algorithme de classification. Il apparait donc indispensable de mettre au point un système de hiérarchisation des variants robustes permettant d'identifier de manière spécifique ces quelques mutations d'intérêt.

De nombreux algorithmes d'apprentissage existent pour le problème de classification. Dans un souci de performance pure, nous avons choisi un modèle d'ensemble, LightGBM (1), combinant la prédiction de plusieurs estimateurs (des arbres de décision) construits grâce à une stratégie de boosting (chaque nouvel estimateur est construit sur la 'pseudo-erreur' des précédents estimateurs). LightGBM a pour principaux avantages d'être hautement paramétrable et performant. Dans un souci de reproductibilité, nous avons réalisé une recherche des hyperparamètres du modèle en se basant sur un jeu de données d'apprentissage composé de 80% de nos variants. Pour cela nous avons eu recours à un algorithme d'optimisation séquentielle de modèle (2) permettant au classifieur de converger vers la meilleure solution par l'évaluation progressive du score F1 moyen de validation croisée à 5 folds. Les hyperparamètres que nous avons considérés ont pour grande majorité trait à la complexité du modèle : par exemple le nombre d'arbres ou leur taille maximum.

A l'issue de l'étape de recherche d'hyperparamètres, le modèle final est entraîné sur la totalité du jeu d'apprentissage, puis validé sur le jeu de test. Ainsi, sur les 51 variants positifs à retrouver, on dénombre 4 faux négatifs. De même, sur les 1933 variants négatifs, 9 sont détectés à tort comme positifs. Le score F1 obtenus est de 0.851 et l'aire sous la courbe (AUC) ROC est de 0.952. Ces résultats témoignent d'une bonne séparation des variants d'intérêts grâce au seuil natif de l'algorithme (probabilité de la classe positive supérieure à 0.5).

Nous avons ensuite tenté de reproduire cette approche sur un jeu de données différent en collectant les données publiques générées par Cohen et al.(3) correspondant à des séquençages d'ADN plasmatique de patients porteurs de différents types de tumeurs solides. Sur ces données, les auteurs avaient identifié essentiellement des mutations sous-clonales. Ainsi, les variants obtenus avaient des VAFs de moyenne et de variance très faibles qui ont nécessités l'utilisation d'autres variant callers plus adaptés à ce genre de problème. De plus, la proportion de variants positifs est bien plus faible que précédemment (environ 0.2% de positifs pour ce jeu de données contre 2% pour le précédent). Ces problématiques rendent la séparation des classes bien plus complexe. Ainsi, nos premières expérimentations indiquent des résultats intéressants mais moins favorables que pour le précédent jeu de données : le score F1 obtenu est de 0.5 et l'AUC ROC est de 0.871. Néanmoins, quelques pistes sont envisagées pour améliorer ces résultats comme le suréchantillonnage synthétique (4) pour pallier la sous-représentation des positifs.

Conclusion

Cette étude nous a permis de mettre en place un modèle statistique permettant de réaliser des hiérarchisations de variants dans deux contextes biologiques différents. Comme nous l'avons noté précédemment, les caractéristiques des variants d'intérêt changent selon le type de problème : ADN tumoral ou plasmatique qui implique par exemple des choix de variant callers différents. Or, jusqu'à présent nous avons entraîné notre modèle sur des jeux de données représentatifs d'un problème spécifique. La capacité de notre modèle à généraliser doit être ainsi mesuré, ce qui fera l'objet de travaux futurs.

Cet outil de classification de variants est en cours d'implémentation dans notre pipeline bioinformatique actuel. Notre intention n'est pas d'exclure automatiquement les variants

prédits négatifs (au risque d'exclure à tort des faux négatifs) mais de produire pour le biologiste un score qui sera complémentaire de son expertise.

References

1. Ke G, Meng Q, Finley T, Wang T, Chen W, Ma W, et al. LightGBM: A Highly Efficient Gradient Boosting Decision Tree. In: Guyon I, Luxburg UV, Bengio S, Wallach H, Fergus R, Vishwanathan S, et al., editors. Adv Neural Inf Process Syst 30 [Internet]. Curran Associates, Inc.; 2017 [cited 2019 Mar 28]. page 3146–3154. Available from: http://papers.nips.cc/paper/6907-lightgbm-a-highly-efficient-gradient-boosting-decision-tree.pdf

2. Bergstra JS, Bardenet R, Bengio Y, Kégl B. Algorithms for Hyper-Parameter Optimization. In: Shawe-Taylor J, Zemel RS, Bartlett PL, Pereira F, Weinberger KQ, editors. Adv Neural Inf Process Syst 24 [Internet]. Curran Associates, Inc.; 2011 [cited 2019 Mar 28]. page 2546–2554. Available from: http://papers.nips.cc/paper/4443-algorithms-for-hyper-parameter-optimization.pdf

3. Cohen JD, Li L, Wang Y, Thoburn C, Afsari B, Danilova L, et al. Detection and localization of surgically resectable cancers with a multi-analyte blood test. Science. 2018;359:926–30.

4. Chawla NV, Bowyer KW, Hall LO, Kegelmeyer WP. SMOTE: Synthetic Minority Oversampling Technique. J Artif Intell Res. 2002;16:321–57.

Structuration et consolidation de résultats d'analyses de RNAseq et Polymorphisme

Thomas GARCIA[1], Ludovic LEGRAND[1], Jérôme GOUZY[1] and Sébastien CARRERE[1]

[1] LIPM, Université de Toulouse, INRA, CNRS, Castanet-Tolosan, France

Corresponding Author: sebastien.carrere@inra.fr

L'accumulation de résultats d'analyses « omiques » au cours du temps et par différents acteurs rend leur utilisation et leur intégration difficile. La légumineuse modèle *Medicago truncatula* est un exemple de production de données de transcriptomiques et de polymorphisme sur plus de 20 ans et ce sur plusieurs versions d'assemblage de génome de référence. Les résultats publiés nécessitent des étapes de transformation et de normalisation afin de pouvoir être intégrés. Pour palier ce problème, nous construisons une solution reposant sur 3 briques logicielles :

1. La *collecte* se fait via une instance de l'ARCHIVE [1] pour structurer les séquences et annotations de référence utilisées, les outils et les métadonnées de l'analyse d'origine ainsi que les résultats d'analyse dans des formats textes standards (vcf, tsv).

2. La *transformation* et la *normalisation* s'appuient sur les métadonnées accompagnants les résultats d'analyses afin d'en extraire les groupes de répétition biologiques et/ou techniques. Les métadonnées sont également utilisées pour calculer des statistiques descriptives qui seront mises à disposition de l'utilisateur pour visualiser rapidement les résultats des analyses disponibles dans l'environnement pour l'espèce d'intérêt. Ainsi, pour les résultats d'analyses de RNAseq, les valeurs de comptage de chaque échantillon sont alors sommées par répétition technique afin d'obtenir une valeur par échantillon et par objet biologique. Ces valeurs sont ensuite normalisées suivant trois méthodes : CPM, TPM et RPKM. Des statistiques descriptives sont calculées pour chaque condition comme contrôle qualité. La transformation des résultats d'analyses de polymorphisme consiste en l'analyse de chaque position polymorphe afin d'en extraire l'impact fonctionnel, sa présence ou non dans chaque échantillon (NoCall) et sa nature (HomRef, HomVar, Het).

3. La *consolidation* de l'ensemble des résultats issus de l'étape précédente s'effectue en intégrant des informations d'annotation fonctionnelle nécessaires à l'interprétation des résultats. Pour cela, un fichier d'annotation du génome de l'organisme, identifié lors de la collecte, est analysé afin d'en extraire des informations pertinentes pour chaque objet biologique (nom, accession, locus_tag, fonction). Comme il peut exister plusieurs versions d'annotation pour un seul génome, la consolidation des résultats d'analyses est réalisée en fonction de la version renseignée dans les métadonnées de l'analyse lors de sa collecte.

Les résultats d'analyses ainsi consolidés viennent alors renforcer un ensemble cohérent et sont mis à disposition au travers d'une interface web, avec pour chaque analyse un ensemble de statistiques descriptives. L'utilisateur peut alors sélectionner spécifiquement les résultats qu'il pense cohérent pour être utilisés dans son analyse en fonction d'un organisme et d'une version d'annotation.

Les données sélectionnées sont mises à disposition en un fichier téléchargeable par l'utilisateur dans un format exploitable via des outils couramment utilisés en laboratoire (R, Excel).

References

[1] https://bbric.toulouse.inra.fr/reference
[2] Pecrix et al., *Whole-genome landscape of Medicago truncatula symbiotic genes.* Nat. Plants, 2018.

Study of sperm epigenetic contribution for the regulation of embryonic gene transcription in early development

Valentin FRANÇOIS - - CAMPION[1] and Jerome JULLIEN[2]

[1] CRTI, INSERM, Université de Nantes, Nantes, France

[2] Wellcome Trust CRUK Gurdon Institute, University of Cambridge, United Kingdom

Corresponding Author: valentin.francois-----campion@univ-nantes.fr

So far, we have provided evidence that the sperm is epigenetically programmed to regulate embryonic development [1]. We have shown that this epigenetic programming is in part related to the regulation of embryonic gene expression at the time of development when the embryo first starts to transcribe genes (Zygotic Gene Activation, ZGA). We and others have shown that histone methylation in sperm is associated with developmentally important genes (Xenopus [1], mouse [2], human [3,4], zebrafish [5]). We have shown for the first time that some part of the genome are found methylated in every sperm cells. We have also shown that some sperm derived methylated histones are maintained in the cells of the developing embryos [1]. Lastly, we and other have obtained indirect evidence that sperm modified histones are required for gene expression in embryos. Indeed, interference with epigenetic marks during the formation of the gametes or after fertilization lead to embryonic gene misregulation [6,7,8]. We showed that the disruption of H3K4me3 and H3K27me3 in sperm alter the embryo development by genes expression misregulation. We find that HOX genes involved in early embryo development are disrupted in case of these epigenetics marks deregulation. We have recently devised experimental strategy that enables epigenetic modification of a mature sperm. We incubate sperm in an oocyte extract that contains a chromatin modifier, the H2A deubiquitylase USP21. In such extract, the sperm nucleus is partially decondensed, allowing chromatin modifier to access directly its chromatin target. We characterize gene expression between haploid embryo generated from control and H2AK119ub depleted sperm (Alignment with HISAT2, RNA-seq analysis with edgeR in progress). We test if genes differentially expressed between embryos generated from control and H2AK119ub depleted sperm correlate to the presence of H2Aub mark around these genes in regulation area (ChIP-seq analysis, alignment with bowtie2 and peak calling with MACS2 in progress). Then, we identify genes whose regulation is likely to be directly regulated by the presence of H2AK119ub. We will have to investigate some meta-analysis whether they also have particular H3K4me3/H3K27me3 marking as well as positioned nucleosome in sperm. These candidates' genes will then be selected for targeted specific epigenome modification of the sperm using dcas9-USP21 fusion protein. The aim is to produce more haploid embryos in the future and evaluate the impact on the expression of the modified genes with our candidates' genes from all analyses.

References

1. Teperek, M. et al. Sperm is epigenetically programmed to regulate gene transcription in embryos. *Genome Res 26*, 1034-1046 (2016).
2. Erkek, S. et al. Molecular determinants of nucleosome retention at CpG-rich sequences in mouse spermatozoa. *Nat Struct Mol Biol* 20, 868-875 (2013).
3. Brykczynska, U. et al. Repressive and active histone methylation mark distinct promoters in human and mouse spermatozoa. *Nat Struct Mol Biol* 17, 679-687 (2010).
4. Hammoud, S.S. et al. Distinctive chromatin in human sperm packages genes for embryo development. *Nature* 460, 473-478 (2009).
5. Wu, S.F., Zhang, H. & Cairns, B.R. Genes for embryo development are packaged in blocks of multivalent chromatin in zebrafish sperm. *Genome Res* 21, 578-589 (2011).
6. Vastenhouw, N.L. et al. Chromatin signature of embryonic pluripotency is established during genome activation. *Nature* 464, 922-926 (2010).
7. Murphy, P.J., Wu, S.F., James, C.R., Wike, C.L. & Cairns, B.R. Placeholder Nucleosomes Underlie Germline-to-Embryo DNA Methylation Reprogramming. *Cell* 172, 993-1006 e1013 (2018).
8. Siklenka, K. et al. Disruption of histone methylation in developing sperm impairs offspring health transgenerationally. *Science* (2015).

Supervised contact prediction in proteins

Maureen Muscat[1], Giancarlo Croce[1], Edoardo Sarti[1] and Martin Weigt[1]
Laboratory of Computational and Quantitative Biology, Sorbonne Université 4 Place Jussieu, 75005, Paris, France

Corresponding author: maureen.muscat@upmc.fr

1 Abstract

Proteins are the major work horses of the cell. Being part of all essential biological processes, they have catalytic, structural, transport, regulatory and many other functions. The prediction of the 3D structure of protein is a key research area in bioinformatics since what a protein can do depend on its unique shape.

Thanks to the rapid reduction in the cost of genetic sequencing and the fast growth of biological sequence databases, different approaches that rely on genomic data have become increasingly popular in the last few years.

A particularly successful method is Direct Coupling Analysis (DCA [1,2]), an unsupervised machine learning approach, which exploits the co-evolutionary signal contained in the multiple sequence alignments to predict residues which are in contact (Fig. 1).

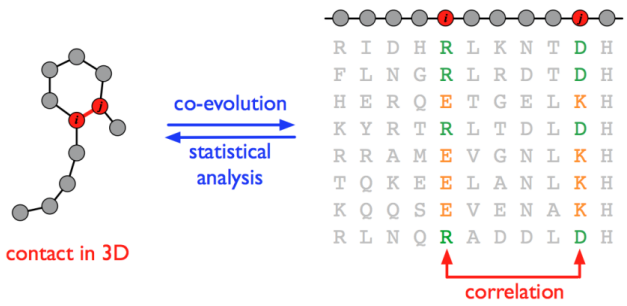


Fig. 1. *Left*: a protein structure with 2 amino acids in contact. *Right*: the corresponding amino acids coevolving in the MSA [1]

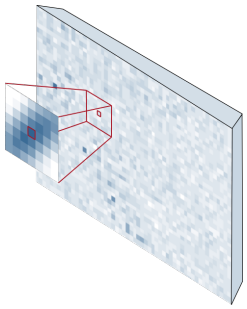


Fig. 2. Our approach: applying "filters", learned from experimental structures, on the DCA predicted contact map.

We aim to improve DCA by using a supervised approach: we observe that the contact maps of proteins are not random matrices, in fact we find different patterns related to the secondary structure. Our idea is to learn these patterns from experimental structures of proteins and then use them as "filters" to be applied on the DCA predicted contact map (Fig. 2).

Our method, despite its simplicity and interpretability, has been proved to greatly outperform the performace of DCA and, in the future, we plan to develop a version oriented to the prediction of interaction interfaces (interaction domain-domain and protein-protein).

References

[1] Faruck Morcos, Andrea Pagnani, Bryan Lunt, Arianna Bertolino, Debora S Marks, Chris Sander, Riccardo Zecchina, José N Onuchic, Terence Hwa, and Martin Weigt. Direct-coupling analysis of residue coevolution captures native contacts across many protein families. *Proceedings of the National Academy of Sciences*, 108(49):E1293–E1301, 2011.

[2] Magnus Ekeberg, Cecilia Lövkvist, Yueheng Lan, Martin Weigt, and Erik Aurell. Improved contact prediction in proteins: Using pseudolikelihoods to infer Potts models. *Physical Review E*, 87(1):12707, January 2013.

Symmetries of the hypercube : a tool for regulatory networks analysis

Jean Fabre-Monplaisir, Brigitte Mossé and Élisabeth Remy
Aix Marseille Univ, CNRS, Centrale Marseille, I2M, Marseille, France

Corresponding author: jean.fabre-monplaisir@ens.fr, elisabeth.remy@univ-amu.fr

In the context of genetic regulatory networks, dynamics of a cell can be modeled by a boolean function S: a map from the hypercube $\{0,1\}^n$ to itself, with n the number of genes [1,2]. Each boolean variable x_i states for the discrete expression of gene g_i (present/absent), and boolean vectors $x \in \{0,1\}^n$ are the states of the system. From S, we can compute the corresponding regulatory graph $\mathcal{RG}(S)$ (oriented and signed), consisting in all activations and inhibitions (edges with sign $+$, resp. $-$) between genes (nodes). If a gene has at least one regulator, we describe its requirements to be activated by logical rules involving all its regulators [1,3].

Properties of S can be interpreted as biological features (e.g. multistability and cellular differentiation, cyclical attractors and homeostasis) [1].

Given S a boolean dynamics and f a symmetry of the hypercube, we consider the conjugated dynamics $\phi_f(S) = f \circ S \circ f^{-1}$ [4,5]. Clearly, $\phi_f(S)$ conserves the dynamical properties of S [6]; we compare their respective regulatory graphs $\mathcal{RG}(S)$ and $\mathcal{RG}(\phi_f(S))$, and their logical rules.

The regulatory graphs of two conjugated dynamics S and $\phi_f(S)$ are similar: our study proves that they have the same topology (nodes are renumbered); edges may switch their signs, but signs of circuits remain unchanged. Their logical rules may also be modified; for example, logical operators OR and AND may be interchanged between two conjugated dynamics, but not OR and XOR.

The set of symmetries of the hypercube defines classes of boolean dynamics, gathering all the conjugates $\phi_f(S)$ of a given boolean dynamics S, in other words gathering all the isometric dynamics. Thus, we aim at classify the set of boolean dynamics on the basis of those isometries, and emphasize their common features through regulatory graphs and logical rules. We can then restrict the dynamical analysis of all the boolean functions to one representant per class.

As an illustration, we study boolean dynamics of well-known motifs - isolated circuits, chorded circuits and flower graphs - through the choice of an appropriate representant.

This leads to a comparison between isometric and isomorphic dynamics (in terms of directed graphs on the hypercube), both natural tools for classification. Clearly, isometric dynamics are isomorphic, and under some restrictions, isometric and isomorphic are equivalent.

References

[1] René Thomas. On the relation between the logical structure of systems and their ability to generate multiple steady states or sustained oscillations. *Springer Series in Synergetics*, 9:180–193, 1981.

[2] Claudine Chaouiya, Élisabeth Remy, Brigitte Mossé, and Denis Thieffry. Qualitative Analysis of Regulatory Graphs : A Computational Tool Based on a Discrete Formal Framework. In *Lecture Notes in Control and Information Science*, volume 294, pages 119–126, 2003.

[3] Stuart Kauffman. The ensemble approach to understand genetic regulatory networks. *Physica A: Statistical Mechanics and its Applications*, 340(4):733–740, sep 2004.

[4] David Slepian. On the number of symmetry types of boolean functions of n variables. *Canadian Journal of Mathematics*, 5:185–193, 1953.

[5] Leon Glass. Classification of biological networks by their qualitative dynamics. *Journal of theoretical biology*, 54(1):85–107, oct 1975.

[6] Elisa Tonello, Etienne Farcot, and Claudine Chaouiya. Local negative circuits and cyclic attractors in boolean networks with at most five components. *SIAM Journal on Applied Dynamical Systems*, 18(1):68–79, 2019.

Séquençage d'ADN natif dédié à l'étude du microbiome sur le MinION ® : retour d'expérience de la paillasse à l'assignation taxonomique

Ségolène Diry[1], Alban Ott[1], Léo D'agata[1], Eric Ginoux[1], and Virginie Chesnais*[1]

[1]LifeSoft – Entreprise privée – France

Résumé

Mots clés: Native whole genome sequencing, Nanopore, Microbiome

Introduction

Le microbiome est un acteur majeur dans la régulation de la physiologie chez l'homme. L'exploration du microbiome d'un échantillon est souvent réalisé grâce au séquençage de tout ou partie du gène codant pour l'ARN 16S qui est un gène conservé au sein des bactéries. Une autre approche plus coûteuse consiste à séquencer tout l'ADN d'un échantillon (whole genome shotgun sequencing) et permet d'accéder à l'ensemble du microbiome, incluant donc les bactéries mais aussi les microorganismes eucaryotes unicellulaires (ex : levures) et les virus/phages. Bien que plus coûteuse, elle apporte une meilleure description de la diversité d'un échantillon tout en fournissant d'autres informations telles que l'analyse fonctionnelle (1). Actuellement, une limite majeure de ces méthodes est la taille restreinte de la région du 16S séquencé qui découle de la taille des reads obtenus après séquençage. Une deuxième limite est la nécessité d'amplifier l'ADN par PCR en amont du séquençage. En effet, cette étape peut générer des biais dans l'abondance de certaines espèces notamment du fait du nombre variable de copies du gène 16S dans les différents génomes de micro-organismes. Le développement de la technologie Nanopore® permet de répondre à ces limites en générant des reads suffisamment longs pour réaliser des assignations taxonomiques plus précises d'une part et en séquençant directement l'ADN natif d'un échantillon d'autre part. Des études ont déjà montré que le séquençage entier du gène codant pour l'ARN 16S grâce à cette technologie permettait d'améliorer la classification taxonomique d'un échantillon (2). Finalement, l'un des enjeux de ce type d'analyse est la caractérisation du microbiome d'un échantillon en un temps limité, afin d'améliorer la prise en charge des patients dans le cas d'infections bactériennes par exemple.

Afin de répondre à ces différentes problématiques nous avons utilisé la technologie Nanopore® pour séquencer l'ADN natif complet en long-read d'un échantillon. Par cette approche, nous cherchons à mettre au point, de façon routinière et rapide, la caractérisation du microbiome d'un échantillon par séquençage. Pour cela nous avons réalisé plusieurs tests en faisant varier aussi bien les étapes de préparation des échantillons (méthodes d'extraction d'ADN...) que les étapes bioinformatiques (base calling, assignation taxonomique...).

*Intervenant

Résultats

Nous avons réalisé des séquençages d'ADN bactérien sur MinION® grâce au kit SQK-RBK004. En amont du séquençage nous avons comparé différentes méthodes d'extraction de l'ADN afin d'évaluer leur incidence sur la taille et la qualité des reads produits après séquençage. A partir des différents jeux de données générés, nous avons testé différents outils d'analyses depuis les étapes de base calling des données brutes jusqu'à l'assignation taxonomique des reads. Afin de choisir le meilleur pipeline d'analyse possible, chaque outil a été également évalué sur des données in-silico, c'est-à-dire des données de séquençage Nanopore® générées artificiellement et mimant différentes configurations, afin de les valider et de déterminer les paramètres optimaux pour nos analyses.

Dans un premier temps nous avons compareé 3 méthodes d'extraction d'ADN sur différents types d'échantillons afin d'évaluer leur impact sur les résultats de séquençage. Nous avons également comparé les résultats obtenus par 3 approches de base calling et de démultiplexage. Dans tous les cas nous avons observeé la présence d'un taux important (10% à 40%) de reads qui ne sont assignés à aucun barcode quel que soit les outils d'analyse utilisés. Nous avons également pu observer une différence dans la taille médiane des reads générés selon les protocoles d'extraction utilisé, de même qu'une variabilité de la proportion de reads ayant une qualité supérieure à 7. Ces observations suggèrent l'importance du choix de la méthode d'extraction de l'ADN : plus les fragments d'ADN obtenus après extraction sont grands, plus les reads obtenus après séquençage seront de taille et de qualité suffisante pour être correctement assignés. Enfin, en comparant les reads assignés à un barcode, nous avons pu mettre en évidence que seuls 87% à 92% des reads étaient assignés de la même façon selon la méthodologie choisie. Ces reads assignés différemment selon les outils semblent majoritairement être des reads de plus faible qualité confirmant l'importance de filtrer les reads de mauvaise qualité pour la suite des analyses afin de ne pas générer de biais dans les résultats dû à la mauvaise assignation d'un read.

Nous avons ensuite comparé deux approches différentes pour réaliser l'assignation taxonomique : l'alignement direct des reads sur différentes bases de données et l'alignement de contig générés par assemblage de novo. Pour évaluer ces deux approches nous avons séquencé des échantillons synthétiques composés de 1 à 4 espèces bactériennes différentes. L'alignement direct des reads sur la base de données de génomes bactériens HMRGD permet l'alignement de 93% à 98% des reads sur les espèces attendues et de près de 99% des reads sur des bactéries du même genre que celles attendues. Pour les alignements observés sur des espèces très différentes que celles attendues nous avons pu mettre en évidence une augmentation significative du taux d'insertion et de délétion de ces reads suggérant un alignement de qualité moindre pour ces espèces. L'approche par assemblage de novo nous a permis de générer des contigs s'alignant pour 99% d'entre eux sur les espèces attendues. Cette seconde stratégie semble donc donner des résultats plus spécifiques mais nécessite une profondeur de séquençage suffisante afin de permettre la création de contigs pour chaque espèce attendue. On peut en effet supposer une sensibilité plus faible de cette approche pour l'identification des espèces minoritaires présentes dans un échantillon.

Pour finir nous avons validé notre méthodologie sur des échantillons réels : des prélèvements du microbiote de peau présent sur les joues de plusieurs sujets sains. L'alignement des reads sur des bases de données de bactéries, d'archeae et de champignons nous a permis de visualiser la diversité du microbiote de la peau chez ces sujets. Pour chaque échantillon nous avons dénombrer en moyenne 87% de bactéries, 8% de champignons et 4% d'archae. De manière intéressante, nous avons observé une bonne corrélation entre les résultats obtenus pour les prélèvements réalisés sur la joue droite et la joue gauche de chaque sujet (r spearman moyen : 0.75). En se basant sur l'indice de Shannon, des tests de raréfaction nous ont également permis de montrer que la génération d'un minimum de 800 reads par échantillon semblaient suffisants pour refléter toute la diversité d'un échantillon.

Conclusion
Au final, cette étude nous a permis de comparer différentes stratégies de base calling et

de démultiplexage. Nous avons également pu évaluer l'importance des filtres qualité en amont de l'analyse et pu déterminer la meilleure approche pour l'assignation taxonomique d'un échantillon. De même, nous avons mis en évidence l'importance du choix de la technique d'extraction de l'ADN qui a un impact important sur la longueur et la quantité de reads obtenus par le séquenceur et la qualité des résultats en fin d'analyse. Nos différentes mises au point ont permis d'aboutir à un workflow permettant d'obtenir la classification taxonomique d'un échantillon en journée depuis le prélèvement biologique jusqu'à la génération du rapport taxonomique. En plus de la rapidité de ce protocole, le séquençage de l'ADN complet d'un échantillon grâce à la technologie Nanopore® ouvre de multiples opportunités pour la caractérisation plus fine des échantillons, notamment l'identification d'archeae et de champignons à partir des mêmes résultats de séquençage ainsi que la possibilité de réaliser des analyses fonctionnelles telles que la prédiction d'ORF.

The ClermonTyper: an easy-to-use and accurate *in silico* tool for *Escherichia* genus strain phylotyping

Bénédicte Condamine[1], Antoine Bridier-Nahmias[1], Hervé Le Nagard[1], Johann Beghain[1], Erick Denamur[1,2], and Olivier Clermont[1]

[1] IAME UMR1137 INSERM, Université Paris Diderot Sorbonne Paris Cité, F-75018 Paris, France

[2] Assistance Publique-Hôpitaux de Paris, Hôpital Bichat, Laboratoire de Génétique Moléculaire, F-75018 Paris, France

Corresponding Author: benedicte.condamine@inserm.fr

The ClermonTyping method was recently developed to assign Escherichia strains to a particular species or a phylogroup[1]. Indeed the genus *Escherichia* is composed of *Escherichia albertii*, *E. fergusonii*, five cryptic *Escherichia* clades and *E. coli sensu stricto.* The later was until now divided into seven main phylogroups termed A, B1, B2, C, D, E and F. The ClermonTyping was inspired [2] by the *in vitro multiplex* PCR assays developped in the lab that allow the identification of most of these species/phylogroups. The ClermonTyper checks for the presence of primer sequences of the genes used for the *in vitro* PCR in assembled genome sequences. Then, a decision algorithm is used to attribute a species/phylogroup affiliation to the query sequence. The result is then confronted with a whole genome comparison method based on Mash [3]. It was recently noted that some strains wrongly assigned to the group F were at an intermediate distance between the F and the B2 groups [4]. These strain form the group G and we designed an *in vitro* PCR assay and its *in silico* counterpart to classify strains from this group. We are now updating the ClermonTyper and its to enable identification of this new phylogroup.

Keywords

Escherichia coli, phylogroups, ClermonTyper

References

[1] Beghain Johann, Bridier-Nahmias Antoine, Le Nagard Hervé, Denamur Erick and Clermont Olivier: ClermonTyping: an easy-to-use and accurate in silico method for Escherichia genus strain phylotyping. *Microbial Genomics,* 2018

[2] Clermont Olivier, Julia K. Christenson, Erick Denamur ans David M. Gordon: The Clermont Escherichia coli phylo-typing method revisited: improvement of specificity and detection of new phylo-groups. *Environmental microbiology reports,* 2013

[3] Brian D. Ondov, Todd J. Treangen, Páll Melsted, Adam B. Mallonee, Nicholas H. Bergman, Sergey Koren, and Adam M. Phillippy: Mash: fast genome and metagenome distance estimation using MinHash. *Genome Biology,* 2016

[4] Teemu Kallonen, Hayley J. Brodrick, Simon R. Harris, Jukka Corander, Nicholas M. Brown, Veronique Martin, Sharon J. Peacock and Julian Parkhill: Systematic longitudinal survey of invasive Escherichia coli in England demonstrates a stable population structure only transiently disturbed by the emergence of ST131. *Genome Research*, 2017

The extra mile of Gene Set Enrichment Analysis: seeing the data

Ahmed Kéceli[1] and Natalia Pietrosemoli[1]
HUB Bioinformatics and Biostatistics - C3BI - Institut Pasteur - USR 3756 CNRS, 25-28 Rue du Dr Roux, 75015, Paris, France

Corresponding author: natalia.pietrosemoli@pasteur.fr

1 Abstract

Gene Set Enrichment Analysis (GSEA) is a powerful approach for interpreting transcriptomic and proteomic data based on the functional annotation of genes. Such methods are useful for identifying enriched gene sets, i.e.,groups of related genes involved in the same molecular pathways, biological processes, cellular compartments, or in other common biological features [1]. The resulting enriched gene sets may provide information on the biological functions and processes associated with specific experimental conditions: cell populations, response to treatment, etc. GSEA methods, thus, provide an easy transition of the classical gene-space based differential analysis (DA), to a complementary gene-set-space based analysis that, unlike DA, takes into account the interactions among genes.

In the last years, a profusion of methods have emerged, reflecting their popularity in OMIC data analysis. Indeed, GSEA methods reduce the analysis complexity to a few hundreds of gene sets (or molecular pathways), identify the active gene sets/pathways differing between the conditions and increase the explanatory power as compared to a simple list of differentially expressed genes (DEGs) [2]. Even if GSEA tools vary in their statistical implementation and in the gene set collections they rely on, they share, overall, the same output format. It basically consists of i) a list of gene sets enriched in a specific biological condition, ii) the number of genes in the dataset belonging to each gene set, and iii) the p-value associated with the test statistic. This format does not differ much from that of the list of DEGs, and indeed they both fail to reveal the relationship among the listed gene-sets or DEGs. Moreover, except for a few exceptions, GSEA methods do not provide any graphical representation of the results. Visualization of results, however, is a crucial step of data analysis: a good visualization makes possible to present the biological complexity of the results and provides a context to gain insights on them.

We propose an interactive pipeline to generate an ensemble of GSEA visualizations derived from the results of the EGSEA R package [3], including i) heatmaps comparing the level of enrichment of each gene set, ii) boxplots exploring the distribution of the fold changes in the expression genes in the enriched gene sets, and iii) arc-plots and graph representations showing the amount of shared genes among the gene sets. We propose also individual heatmaps of the expression of genes belonging to a specific gene-set. Offering the combination of gene-set level and gene level information in a single analysis together with ad-hoc visualization options will provide a more intuitive and easier interpretation of the results while allowing to explore the results at different granularity levels leading to a broader biological interpretation.

2 Key words

differentially expressed genes, functional analysis, molecular pathway analysis, RNA-Seq, transcriptomics

References

[1] Adi L. Tarca, Gaurav Bhatti, and Roberto Romero. A Comparison of Gene Set Analysis Methods in Terms of Sensitivity, Prioritization and Specificity. *PLOS ONE*, 8(11):e79217, November 2013.

[2] Purvesh Khatri, Marina Sirota, and Atul J. Butte. Ten Years of Pathway Analysis: Current Approaches and Outstanding Challenges. *PLoS Computational Biology*, 8(2):e1002375, February 2012.

[3] Monther Alhamdoosh, Milica Ng, Nicholas J. Wilson, Julie M. Sheridan, Huy Huynh, Michael J. Wilson, and Matthew E. Ritchie. Combining multiple tools outperforms individual methods in gene set enrichment analyses. *Bioinformatics*, page btw623, September 2016.

The limit of cell specification concept: a lesson from scRNA-Seq on early human development.

Dimitri Meistermann[1,2,4], Sophie Loubersac[1,3], Arnaud Reignier[1,3],
Valentin Francois - - Campion[1,2], Thomas Fréour[1,3], Jérémie Bourdon[4], Laurent David[1,2]

[1] CRTI, INSERM, UNIV Nantes, Nantes, France;
[2] ITUN, CHU Nantes, Nantes, France;
[3] Service de Biologie de la Reproduction, CHU Nantes, Nantes, France;
[4] LS2N, CNRS, UNIV Nantes, Nantes, France;

Corresponding Author: dimitri.meistermann@univ-nantes.fr

Recent technological advances such as single-cell RNAseq have allowed an unprecedented access into processes orchestrating human preimplantation development **[1, 2]**. However, the sequence of events which occur during human preimplantation development are still unknown. In particular, timing of the very first human lineage specification remains elusive. During this event, the morula cells are can acquire two fates: the trophectoderm that will give rise the placenta and inner cell mass that will give rise the fetus. We present a human preimplantation development model based on transcriptomic pseudotime modelling of four scRNAseq dataset, biologically validated by spatial information and precise time-lapse staging. In contrast to mouse **[3]**, we show that trophectoderm / inner cell mass lineage specification in human is only detectable at the transcriptomic level at the blastocyst stage, just prior to expansion. By studying this delay, we show that cellular specification is a time window that begins with the establishment of cellular junctions, which polarize the embryo. These are the first factors that discriminates the two cell fates. The cell specification ends with the divergence of transcriptome profiles. For identifying the precise timings of this divergence, we have coupled the pseudotime modelling from Monocle2 **[4]** with several other tools. First, we performed an estimation of RNA velocity with velocyto **[5]**. This tool can retrieve the genes that are going to be down or upregulated in each cell, by processing the intron data that are contained into scRNAseq reads. We used WGCNA **[6]** for describing the waves of genes that paces human preimplantation development. By combining these tools, we found novel markers, validated by immunofluorescences. Their expression profile enables a precise staging of human preimplantation embryos, such as IFI16 which highlights establishment of epiblast and NR2F2 which appears at the transition from specified to mature trophectoderm. Strikingly, mature trophectoderm cells arise from the polar side, just after specification, supporting a model of polar trophectoderm cells driving trophectoderm maturation. Altogether, our study unravels the first lineage specification event in the human embryo and provides a browsable resource, based on d3.js, for mapping spatio-temporal events underlying human lineage specification.

References

[1] L. Yan *et al.*, « Single-cell RNA-Seq profiling of human preimplantation embryos and embryonic stem cells », *Nature Structural and Molecular Biology*, vol. 20, nº 9, p. 1131, sept. 2013.
[2] S. Petropoulos *et al.*, « Single-Cell RNA-Seq Reveals Lineage and X Chromosome Dynamics in Human Preimplantation Embryos », *Cell*, vol. 165, nº 4, p. 1012-1026, mai 2016.
[3] E. Posfai *et al.*, « Position- and Hippo signaling-dependent plasticity during lineage segregation in the early mouse embryo », *eLife*, vol. 6.
[4] C. Trapnell *et al.*, « The dynamics and regulators of cell fate decisions are revealed by pseudotemporal ordering of single cells », *Nature Biotechnology*, vol. 32, nº 4, p. 381-386, mars 2014.
[5] G. La Manno *et al.*, « RNA velocity of single cells », *Nature*, vol. 560, nº 7719, p. 494-498, août 2018.
[6] P. Langfelder et S. Horvath, « WGCNA: an R package for weighted correlation network analysis », *BMC Bioinformatics*, vol. 9, p. 559, 2008.

The Migale bioinformatics platform

Valentin LOUX[1], Sam AH LONE[1], Hélène CHIAPELLO[1], David CHRISTIANY[1], Sandra DÉROZIER[1], Olivier INIZAN[1], Véronique MARTIN[1], Mahendra MARIADASSOU[1], Cédric MIDOUX[1], Olivier RUÉ[1], Sophie SCHBATH[1], Valérie VIDAL[1]

[1] MaIAGE, INRA, Université Paris-Saclay, Domaine de Vilvert, 78350, Jouy-en-Josas, France

Corresponding Author: valentin.loux@inra.fr

The Migale bioinformatics platform is a team of the INRA's MaIAGE research unit (Applied Mathematics and Computer Science, from Genome to the Environment).It has been existing since 2003 and is intended to provide services to the life sciences community.

The Migale platform offers four types of services:

- an open infrastructure dedicated to life sciences data processing,
- dissemination of expertise in bioinformatics,
- design and development of bioinformatics applications,
- data analysis.

Migale is part of the French Institute of Bioinformatics (IFB) and France Génomique projects. It has ISO 9001:2015 certification and has been labelled ISC ("Infrastructure Scientifique Collective") by INRA.

The poster will illustrate the platform's missions and offered services with examples chosen from recent achievements.

http://migale.jouy.inra.fr

The relationship between gene co-expression network connectivity and phenotypic prediction sheds light at the core of the omnigenic theory

Aurélien CHATEIGNER[1], Marie-Claude LESAGE-DESCAUSES[1], Odile ROGIER[1], Véronique JORGE[1], Jean-Charles LEPLÉ[2], Véronique BRUNAUD[3,4], Christine PAYSANT-LE ROUX[3,4], Ludivine SOUBIGOU-TACONNAT[3,4], Marie-Laure MARTIN-MAGNIETTE[3,4,5], Leopoldo SANCHEZ[1], Vincent SEGURA[1]

[1] BioForA, INRA, ONF, 2163, avenue de la pomme de pin, F-45075 Orléans Cedex 2, France
[2] BIOGECO, INRA, Univ. Bordeaux, Cestas, France
[3] Institute of Plant Sciences Paris-Saclay (IPS2), CNRS, INRA, Université Paris-Sud, Université d'Evry, Université Paris-Saclay, Bâtiment 630, Plateau de Moulon, Gif sur Yvette, France
[4] Institute of Plant Sciences Paris-Saclay (IPS2), CNRS, INRA Université Paris-Diderot, Sorbonne Paris-Cité, Bâtiment 630, Plateau de Moulon, Gif sur Yvette, France
[5] MIA-Paris, AgroParisTech, INRA, Paris, France

Corresponding Author: aurelien.chateigner@inra.fr

Recent literature on the differential role of genes within networks [1, 2], including the omnigenic model [3, 4], distinguishes core from peripheral genes in the layout underlying phenotypes. Cores are typically few, each of them highly contributes to phenotypic variation, but because they are so few, they altogether only explain a small part of trait heritability. In contrast, peripherals, each of small influence, are so numerous that they finally lead phenotypic variation.

We collected and sequenced RNA from 459 European black poplars [5] and built co-expression networks to define core and peripheral genes as the most and least connected ones. We computed the role of each of these gene sets in the prediction of phenotypes and showed that cores contribute additively to phenotypes, consistent with a downstream position in a biological cascade, while peripherals interact to influence phenotypes, consistent with an upstream position. Quantitative and population genetics analyses further showed that cores are more expressed than peripherals but they tend to vary less and to be more differentiated between populations suggesting that they are more constrained by natural selection.

Our work is the first attempt to integrate core and peripheral terminologies from co-expression networks and omnigenic theory. In the end we showed, that there seems to be a strong overlap between them, with core genes from co-expression networks likely being a mixture of integrative hubs with a direct effect on phenotype in agreement with the omnigenic theory, and master regulators, which control the overall metabolic flow towards the phenotype.

References

1. Emily B Josephs *et al*. The Relationship between Selection, Network Connectivity, and Regulatory Variation within a Population of Capsella grandiflora. *Genome Biology and Evolution*, (9):1099–1109, 2017.
2. Niklas Mähler *et al*. Gene co-expression network connectivity is an important determinant of selective constraint. *PLoS genetics*, (13):e1006402, 2017.
3. Evan A Boyle *et al*. An Expanded View of Complex Traits: From Polygenic to Omnigenic. *Cell*, (169):1177–1186, 2017.
4. Xuanyao Liu *et al*. Trans effects on gene expression can drive omnigenic inheritance. *bioRxiv*, 425108, 2018.
5. Mesfin Nigussie Gebreselassie *et al*. Near-infrared spectroscopy enables the genetic analysis of chemical properties in a large set of wood samples from Populus nigra (L.) natural populations. *Industrial Crops and Products*, (107):159–171, 2017.

The role of the LNR domain-containing protein explosion in Oithona nana male differentiation (Crustacea; Cyclopoida).

Kevin Sugier*[1], Laurie Bertrand[2], Soheib Kerbache[2], Karine Labadie[2], Laso-Jadart Romuald[3], Nathalie Martins[2], Céline Orvain[2], Emmanuelle Petit[2], Julie Poulain[2], Patrick Wincker[1], Jean-Louis Jamet[4], Adriana Alberti[1], and Mohammed-Amin Madoui[1]

[1]Génomique Métabolique, Genoscope, Institut François Jacob, CEA, CNRS, Univ Evry, Université Paris-Saclay – CEA, CNRS, Université Paris-Saclay – France

[2]Commissariat à l'Energie Atomique (CEA), Institut François Jacob, Genoscope – CEA, Genoscope – France

[3]Génomique Métabolique, Genoscope, Institut François Jacob, CEA, CNRS, Univ Evry, Université Paris-Saclay – CEA, CNRS, Université Paris-Saclay – France

[4]Université de Toulon, Aix Marseille Universités, CNRS/INSU/IRD, Mediterranean Institute of Oceanography MIO UM 110, CS 60584 – CNRS : UMR110, Mediterranean Institut of Oceanography – France

Résumé

Copepods are small planktonic crustacean and the most abundant metazoan on Earth. They play an essential role in the marine trophic web and biogeochemical cycles. The genus *Oithona* is described as iteroparous, cosmopolite and having the highest numerical density. The *Oithona* male paradox obliged it to alternate feeding (immobile) and mating (mobile) phases, but the molecular basis of this paradox is unknown. Therefore, we investigated this sexual dimorphism at the molecular level through genomic, transcriptomic and protein-protein interaction (PPI) analyses.
The *O. nana* genome comparison to other copepods showed an explosion of Lin-12 Notch Repeat domains-containing proteins coding genes (LDPGs). Among the 75 LDPGs detected, several harboured new protein domain associations including trypsin domains.

Transcriptomic analysis of the five different developmental stages showed, in males, an enrichment of LDPGs (24% of total LDPGs) and other genes involved in proteolysis, nervous system regulation and synapse assembly and functioning, and amino acid conversion to glutamate.
From PPI assays, we found one LDPG, up-regulated in juveniles and adults, that forms a trypsin-containing LDP complex and interacted with extracellular matrix (ECM). This suggests energy and amino acids release through the regulation of ECM lysis. Also, one of the males up-regulated LDPG, under selection, plays a role in the regulation of neurogenesis. This suggests a nervous system dimorphism between females and males.

*Intervenant

The SeCoNeMo approach and its application to ICE annotation in *Firmicutes*

Julie LAO[1,2], Thomas LACROIX[2], Gérard GUÉDON[1], Nathalie LEBLOND-BOURGET[1] and Hélène CHIAPELLO[2]

[1] DynAMic, Université de Lorraine, INRA, 54506, Vandoeuvre-Lès-Nancy, France
[2] MaIAGE, INRA, Université Paris-Saclay, 78350, Jouy-en-Josas, France

Corresponding Authors: julie.lao@inra.fr and helene.chiapello@inra.fr

Mobile genetic elements (MGEs) play a key role in bacterial genome evolution by enabling gene acquisition through horizontal gene transfer. Among these elements, Integrative Conjugative Elements (ICEs) are integrated in the chromosome of their hosts and transferred by the conjugation machinery.

Transfer and maintenance into the recipient cell are the two biological functions indispensable for ICEs. Genes and sequences involved in these functions are physically close on the DNA molecule and are respectively named “integration module” and “conjugation module”.

Two characteristics of ICEs can be enlightened: (i) ICEs evolve rapidly mainly through acquisition, loss and exchange of modules and (ii) ICEs, and other MGEs transferred by conjugation, are frequently integrated in tandem arrays or can be nested MGEs resulting to fuzzy bounds. Consequently, the detection and accurate annotation of ICE bounds is a difficult task that requires dedicated bioinformatics approaches.

So far two bioinformatics approaches allow to automatically detect and annotate ICEs in bacterial genomes:

(i) A pipeline set up by Cury *et al.* that delineates ICEs in bacterial genomes by using the core genes that surrounds them [1,2]. This procedure is based on the detection of the conjugation module using the CONJscan module of the MacSyFinder software [3,4] and needs at least 4 different closely-related genomes to enable ICE annotation. Thus the delineation is sensible to the set of genomes used to compute the core-genome.

(ii) The tool developed by ICEberg team available online and in a standalone version [5]. It searches T4SS-type ICEs and AICEs using a ‘Pattern-based hit co-localization’ method. The method detects “signature proteins” of both conjugation and integration modules with HMM profiles and co-localize the hits. Delineation at the nucleotide level is only possible for a few specific type of ICE.

However both approaches can not detect neither nested nor tandem ICEs which are frequently observed in bacterial genomes. Thus, we have designed a new approach for ICE detection and annotation that includes two main steps:

(i) The detection of “signature proteins” of the integration and conjugation modules of ICEs using the ICE signature database previously described in the ICEFinder approach by Ambroset *et al.* [6].

(ii) The search for ICE boundaries with the SeCoNeMo approach (SEarch of COmbined NEsted MOtifs) based on the type of “signature proteins” that was detected in step (i) and a combination of dedicated rules that allow to detect isolated, nested and tandem ICEs.

In this poster, we will present the first results of this method obtained for the annotation of ICEs in genomes of *Firmicutes*.

References

[1] Cury, J., Touchon, M., & Rocha, E. P. (2017). Integrative and conjugative elements and their hosts: composition, distribution and organization. *Nucleic acids research*, 45(15), 8943-8956.
[2] https://github.com/gem-pasteur/Macsyfinder_models/blob/master/models/Conjugation/Tutorial_ICE.ipynb
[3] Abby, S. S., Néron, B., Ménager, H., Touchon, M., & Rocha, E. P. (2014). MacSyFinder: a program to mine genomes for molecular systems with an application to CRISPR-Cas systems. *PloS one*, 9(10), e110726.
[4] Abby, S. S., Cury, J., Guglielmini, J., Néron, B., Touchon, M., & Rocha, E. P. (2016). Identification of protein secretion systems in bacterial genomes. *Scientific reports*, 6, 23080.
[5] M. Liu, X. Li, Y. Xie, D. Bi, J. Sun, J. Li, C. Tai, Z. Deng, H.Y. Ou (2019) ICEberg 2.0: an updated database of bacterial integrative and conjugative elements. *Nucleic Acids Research*, DOI: 10.1093/nar/gky1123.
[6] Ambroset, C., Coluzzi, C., Guédon, G., Devignes, M. D., Loux, V., Lacroix, T., ... & Leblond-Bourget, N. (2016). New insights into the classification and integration specificity of Streptococcus integrative conjugative elements through extensive genome exploration. *Frontiers in microbiology*, 6, 1483.

The SIRP gene family: widespread conservation in animals, haplotypic polymorphisms in humans and its therapeutic consequences for monoclonal antibody reactivity.

Rémi Guimon [1,2], Fabienne Haspot [1,2], Nicolas Poirier [3], Gilles Blancho [2], Sophie Limou [1,2,4], Pierre-Antoine Gourraud [1,2], Nicolas Vince [1,2]

1 Centre de Recherche en Transplantation et Immunologie UMR 1064, INSERM, Université de Nantes, Nantes, France
2 Institut de Transplantation Urologie Néphrologie (ITUN), CHU Nantes, Nantes, France
3 OSE Immunotherapeutics, Nantes, France
4 Ecole Centrale de Nantes, Nantes, France

Corresponding Author: nicolas.vince@univ-nantes.fr

Signal regulatory proteins (SIRPs) are transmembrane receptors proteins involved in immunological signaling. The human SIRP family counts 5 members (SIRPA, SIRPB1, SIRPB2, SIRPD, SIRPG) [1], but only SIRPA and SIRPG functions are well described. SIRPA and SIRPG bind to CD47, a proximate regulator of multiple cell survival/death pathways. SIRPA haplogroups (V1 and V2) have been found to modify protein conformation, which consequently impacts: 1) CD47 affinity, 2) potential transplantation mismatch with allo-antigens leading to graft rejection, and 3) anti-SIRPA antibody in therapeutic use. However, little is known on population allele frequency, on haplotype structure beyond SIRPA, or on the evolution story of the SIRP family members. We therefore performed a complete phylogenetic analysis first throughout the animal kingdom, then within human populations. Our analysis revealed that SIRP genes are widely conserved among animal species, from fishes to primates, demonstrating a potential key biological role. When focusing in humans, and similarly to V1/V2 SIRPA, we identified two major haplogroups in SIRPG, named G1/G2. The distribution of these haplogroups varied over the globe: V2 has been reported mostly in East Asian populations (frequency of 0.40, while 0.10 in other populations), and G1 has reached a high frequency up to 0.81 in East Asians. This might indicate events of natural selection that we plan to further explore by population differentiation quantification (FST estimates). We also measured extended haplotype homozygosity on SIRPA and SIRPG and we defined haplotype length of each haplogroup describing tendency to long LD for V2 and G2. We built a haplotype network for SIRPA exon 3 haplotypes and could observe a dichotomous haplotype frequency division between Chinese (CHB) in one hand and European and African in the other hand (CEU and YRI). SIRPG haplotype network is less clear as the considered region length is wider than SIRPA (19.5kb vs. 358b), nevertheless, we can observe haplotype restriction between populations with a particularly large diversity in Africans (YRI). Investigating further the evolution history of the SIRP genes might indeed reveal important clues, especially on SIRPA and SIRPG, which could broaden our knowledge on this immune-related family and potentially impact transplant survival and therapeutic monoclonal antibodies affinity.

References

1. van Beek et al. (2005). Signal Regulatory Proteins in the Immune System. *The Journal of Immunology*.

Keywords

Signal-regulatory protein - SIRP - Anthropology - Population Genetics - Phylogeny

Transcript-aware Clustering of Orthologous Exons Shed Light on Alternative Splicing Evolution

Diego Javier Zea[1], Hugues Richard[1] and Elodie Laine[1]
LCQB UMR 7238 CNRS, Institut de Biologie Paris Seine, Sorbonne Université, 7 Quai Saint-Bernard, 75005, Paris, France

Corresponding author: elodie.laine@upmc.fr, hugues.richard@upmc.fr

Abstract *We present ThorAxe, an automated tool to disentangle homology relationships between exons. In particular, it helps to define orthologous exon groups from the set of transcripts of an orthologous gene group. This is of prime importance for the evolutionary analysis of gene families by identifying conserved evolutionary units inside eukaryotic genes. In this study, we applied it to MAPK8 and to a small group of genes with known functional alternative splicing events involving homologous exons.*

Keywords alternative splicing, evolution, homology, orthology, exons

1 Introduction

Most eukaryotic genes are comprised of more than one exon. This can lead to the expression of different protein isoforms, coming from different exon combinations, through a process called alternative splicing (AS). This process allows eukaryotic organisms to increase the size of their proteomes without increasing their genomes. What is more, this modularity on gene organization can lead to evolutionary innovation through exon duplication, deletion and shuffling and through exonization of intronic sequences or intron/exon phase changes.

We are interested in characterize evolutionary innovation related to alternative splicing events at the protein level. In particular, how exon mutations across evolution and alternative splicing events (ASEs) can lead to structural changes that can impact protein function. This is particularly relevant to medicinal chemistry. Indeed AS deregulation has been associated with the development of multiple diseases, particularly with neurodegenerative disorders and cancer [1]. Identifying evolutionary conserved ASEs can help selecting isoforms that could serve as new therapeutic targets.

To study AS evolution, it is necessary to define groups of homologous exons across different species. Previous studies addressing this issue aimed at identifying exon creation and loss [2], comparing evolutionary rates withing exons [3] and determining the size of the universe of exons [4]. The protocols used in those studies relied heavily on human intervention and, to our knowledege, no automated pipeline or tool is available to define homologous exons in the presence of ASEs.

Defining pertinent groups of homologous exons across a potentially large number of more or less distantly related species is a challenging problem. For instance, in the presence of highly similar mutually exclusive exons, the notion of homology is insufficient to distinguish the isoforms containing one or the other exon. We need to achieve a higher resolution by disentangling orthology and paralogy relationships. That means being able to identify groups of exons that got differentiated by speciation (orthologous) from those that have emerged from duplication events (paralogous).

Working with exons at the protein level presents multiple issues. For example, exons sharing genomic coordinates, i.e. overlapping exons, can lead to completely different amino acid sequences because of changes in intron/exon phases. Also, they can share the same protein sequence in the shared region but differ in their extremities. Another issue is exon length. Exons can be too short to allow reliable detection of homology. Finally, the divergence between exons coming from duplication events across different species may be of the same order as the one between orthologous exons.

There are also problems associated with the data itself. It is common to find incomplete transcripts and low-quality sequences. Also, because transcript annotation is biased towards some organisms, the number of exons and transcripts can be underestimated. While for a couple of species there is transcript support level information to base the selection of transcripts in experimental evidence, for other species, it is only possible to decide based on the quality of the annotated transcript.

In this work, we propose *ThorAxe*, an automated tool to identify homologous relationship between exons and distinguish orthologous from paralogous exons, starting from an ensemble of transcript observed in a set of species. *ThorAxe* efficiently addresses most of the issues mentioned above. As a first case study, we used a small group of genes that present known ASEs and display common problems linked to the inference of homology and orthology between exons. From that set of genes, we focus on the c-Jun N-terminal kinases and show how the tool can be used to explain the evolution of their ASEs.

2 Materials & Methods

Transcript and exon data, gene trees and orthology relationships between genes were retrieved from Ensembl [5]. For performance reasons, we cluster exons using the Hobomh I algorithm [6] by doing fast pairwise alignments with the striped Smith Waterman algorithm [7]. Multiple sequence alignments are performed using MAFFT [8] in accurate mode. We used HMMER [9] for creating HMM profiles from the multiple sequence alignments of orthologous exons.

3 Results

We are interested in characterizing the functional and structural impact of ASEs. Hence, we only use exons present in the selected species' transcriptomes. *ThorAxe* automatically downloads transcript data from Ensembl and cleans them up. Specifically, it removes redundancies due to exons and transcripts identical at the protein level and deletes incomplete transcripts and low-quality sequences before homology assignation.

To accurately describe the variability encoded in the input transcripts, we define minimal non-redundant units smaller than exons, which we call subexons. In that way, we can express each exon as a concatenation of one or more subexons. For example, a gene comprising two exons sharing some overlap at the end, with one having an extension at the beginning is going to define two subexons (see Fig. 1). One of the subexons is shared by both exons while the other is going to be specific to one. The use of subexons increases resolution at the time of differentiating transcripts and allows the detection of smaller evolutionary units.

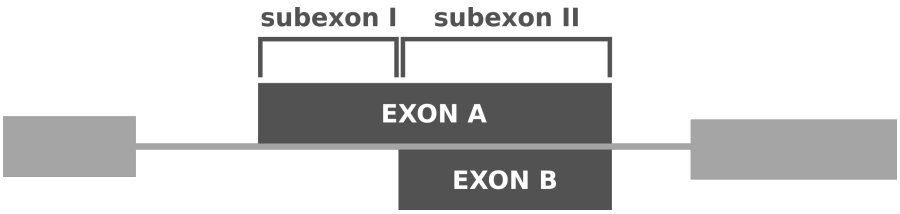


Fig. 1. Overlapped exons A and B defines two subexons if their amino acid sequences are identical.

To define orthologous exon groups we tested at first two different approaches, one based on multiple sequence alignments (MSAs) and one based on pairwise alignments. In the first one, we construct a chimeric protein sequence for each gene as a concatenation of all subexons detected in the expressed transcripts. Subexons are sorted according to their genomic coordinates to reduce the orthology detection problem. Chimeric sequences coming from the different studied species are aligned, and then vertical blocks are defined based on subexon boundaries. These vertical blocks correspond to putative groups of orthologous exons.

The idea behind this first approach is that exons sharing some sequence similarity and that are within the same environment (neighbouring exons) are more likely to be evolutionary related. However, exon shuffling can break this hypothesis by changing the relative location of homologous exons. This approach also fails with subexons pairs sharing low sequence identity because the multiple sequence alignment algorithm forces their alignment.

The approach using pairwise sequence alignments is close to graph-based orthology detection methods [10]. We create a subexon network, with edges joining two subexons if their local alignment is good enough given two thresholds on identity percentage and coverage. This network is used to define clus-

ters of homologous exon. This approach also present problems, in particular, it is unable to distinguish between orthology and paralogy relationships without further processing of the graph. Also, small subexons may artificially link clusters of non-homologous exons. Finally, the thresholds to account for different proteins are hard to set. High thresholds could lead to the loss of homology relationships while low thresholds can create big clusters of non-homologous exons.

To solve those problems, we unified both approaches in the *ThorAxe* pipeline (see Fig. 2). The first clusters of putative homologous exons are defined by pairwise alignments using the Hobohm I algorithm and low thresholds of identity percentage and coverage to ensure that homologous exons belong to the same cluster. This is done before defining subexons to avoid the problem of clustering short sequences. Then, subexons are defined and used to create multiple sequence alignments of chimeric sequences for each cluster to avoid the problem with non-related sequences. Because the clustering is performed at the exon level, to avoid misplaced subexons, a step of refinement is performed over the multiple sequence alignments. Finally, orthologous subexon groups are defined by the vertical blocks present in those alignments.

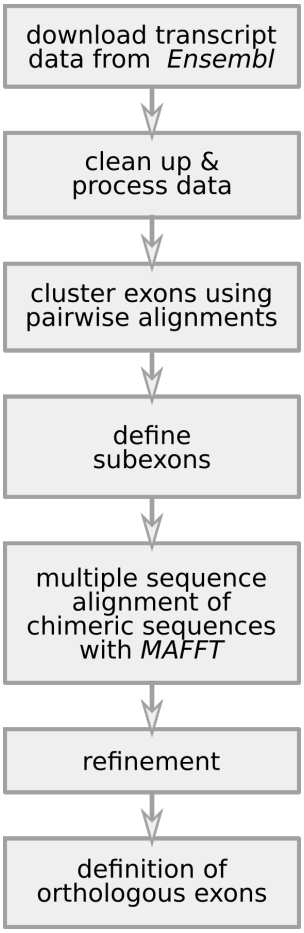


Fig. 2. Main steps in *ThorAxe* pipeline.

ThorAxe pipeline is able to find the mutually exclusive homologous exons pair from MAPK8. *ThorAxe* can identify the oldest one and evidence the possible loss of one of them in some species. It also shows that after the duplication the two exons remains always mutually exclusive. Also, the multiple sequence alignment of both exons allows the comparison of their profile, highlighting high conservation of both with some few conserved differences in both (see Fig. 3).

ThorAxe also output a splice graph where each node represents an orthologous exon group from all the transcripts in the analyzed group of genes, with information about the conservation level (organism fraction) for both, nodes and edges. This helps in the understanding of the evolution of the alternative splice events of the gene group.

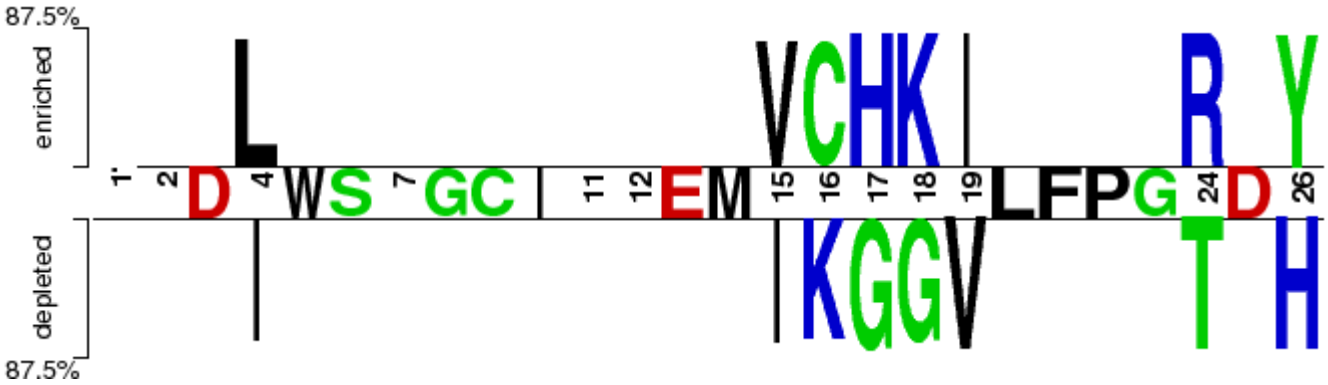


Fig. 3. MAPK8 mutually exclusive homologous exons comparison using *Two Sample Logo* [11].

4 Conclusions

We developed *ThorAxe*, a tool to automatically identify groups of homologous and orthologous exons. Those groups of exons give direct information about the evolutionary conservation of alternative splicing events. Using that data, it will be possible to estimate a date for the appearance of protein isoforms, which can help in the understanding of their biological functions and evolutionary history.

Based on this encouraging preliminary results, we are now applying *ThorAxe* to a curated set of a few dozen proteins and protein families where at least two transcripts with distinct biological functions were experimentally described in order to understand their evolutionnary history.

Acknowledgements

This work was supported by the ANR MASSIV.

References

[1] Amanda J Ward and Thomas A Cooper. The pathobiology of splicing. *The Journal of Pathology: A Journal of the Pathological Society of Great Britain and Ireland*, 220(2):152–163, 2010.

[2] Barmak Modrek and Christopher J Lee. Alternative splicing in the human, mouse and rat genomes is associated with an increased frequency of exon creation and/or loss. *Nature genetics*, 34(2):177, 2003.

[3] Yi Xing and Christopher Lee. Assessing the application of ka/ks ratio test to alternatively spliced exons. *Bioinformatics*, 21(19):3701–3703, 2005.

[4] Robert L Dorit, Lloyd Schoenbach, and Walter Gilbert. How big is the universe of exons? *Science*, 250(4986):1377–1382, 1990.

[5] Daniel R Zerbino, Premanand Achuthan, Wasiu Akanni, M Ridwan Amode, Daniel Barrell, Jyothish Bhai, Konstantinos Billis, Carla Cummins, Astrid Gall, Carlos García Girón, et al. Ensembl 2018. *Nucleic acids research*, 46(D1):D754–D761, 2017.

[6] Uwe Hobohm, Michael Scharf, Reinhard Schneider, and Chris Sander. Selection of representative protein data sets. *Protein Science*, 1(3):409–417, 1992.

[7] Mengyao Zhao, Wan-Ping Lee, Erik P Garrison, and Gabor T Marth. Ssw library: an simd smith-waterman c/c++ library for use in genomic applications. *PloS one*, 8(12):e82138, 2013.

[8] Kazutaka Katoh and Daron M Standley. Mafft multiple sequence alignment software version 7: improvements in performance and usability. *Molecular biology and evolution*, 30(4):772–780, 2013.

[9] Sean R Eddy. Accelerated profile hmm searches. *PLoS computational biology*, 7(10):e1002195, 2011.

[10] Rosa Fernández, Toni Gabaldón, and Christophe Dessimoz. Orthology: definitions, inference, and impact on species phylogeny inference. *arXiv preprint arXiv:1903.04530*, 2019.

[11] Vladimir Vacic, Lilia M Iakoucheva, and Predrag Radivojac. Two sample logo: a graphical representation of the differences between two sets of sequence alignments. *Bioinformatics*, 22(12):1536–1537, 2006.

Transcriptional and functional analyzes of symbiotic coral micro-*algae* in the framework of *Tara* Pacific expedition

Julie Lê-Hoang, Eric Armstrong, Quentin Carradec, Patrick Wincker

Génomique Métabolique, Genoscope, Institut François Jacob, CEA, CNRS, Université d'Évry Val d'Essonne, Université Paris-Saclay

Coral reefs represent a very important ecosystem essential for the life of many marine *species*. They are very well studied in the context of coral bleaching, which increases coral mortality in almost all reefs. This coral mortality is induced by the break of the obligatory symbiosis, between the symbiotic micro-*algae* and the coral host, leading to the expulsion of the micro-*algae* and then the loss of coral coloration (named coral bleaching). This phenomenon is linked to corals under stress conditions, like temperature rising or specific pollutions. In this context, the *Tara* Pacific expedition proposes a vast sampling campaign to study healthy corals at ocean scale and to better understand the adaptation of these organisms to environmental changes. In this context we analyze metatranscriptomic data sequenced from 3 coral *species* (*Millepora platyphylla, Pocillopora meandrina* and *Porites lobata*) sampled around 15 islands by *Tara* consortium in the Pacific Ocean. With these data and bioinformatic methods we are able to (i) Study the different *Symbiodiniaceae species* present in each coral (ii) Looking for genes differentially expressed in the different micro-*algae* between each island and for each coral (iii) Realize functional analyzes in order to observe the transcriptional adaptation of *Symbiodiniaceae*. Among these samples and with different bioinformatic methods we found three different situations where *Symbiodiniaceae* expressions are linked to coral *species*, localization and/or physico-chemical parameters. Overall, biological functions of differentially expressed genes confirm the importance of coral micro-*algae* in the coral reef adaptation in the Pacific Ocean.

Transcriptome analysis to identify co-expressed gene networks as a molecular signature for childhood trauma-related mood disorders.

Amazigh MOKHTARI[1], Bruno ETAIN[2], Ipek YALCIN[3], Cynthia MARIE-CLAIRE[2], El Chérif IBRAHIM[4], Raoul BELZEAUX [4], Pierre-Eric LUTZ[*,3], Andrée DELAHAYE-DURIEZ[*,1,5]

[1]NeuroDiderot - Inserm UMR 1141, Hopital Robert Debré, Paris. [2]Université Paris Diderot, Sorbonne Paris Cité, Inserm, UMR-S1144, Paris, France. [3]Institut des Neurosciences Cellulaires et Intégratives UPR 3212, CNRS, Université de Strasbourg, Fédération de Médecine Translationnelle de Strasbourg, Strasbourg. [4]Université Aix-Marseille, CNRS, Institut de Neurosciences de la Timone, Marseille, France. [5]Université Paris 13, AP–HP, Inserm, Bondy, France. *These authors jointly supervised this work.

Abstract: Bipolar Disorder (BD) and Major Depressive Disorder (MDD) are common and severe psychiatric diseases which can be devastating for the patients, resulting in higher risks of suicide, drug abuse, and shorter life expectancy [1]. Despite the heterogeneous etiology of these illnesses, ranging from genetic predisposition to environmental factors, recent studies showed significant associations between childhood trauma (CT), the severity of symptoms and early onset of BD and MDD [2]. Furthermore, patients who experienced adversity during childhood have been found less responsive to Lithium, the standard mood-stabilizer [3]. Acknowledging that the prognosis of both BD and MDD strongly depends on thymic relapses, understanding their risk factors and underlying mechanisms is, hence, a major challenge to target patients requiring more intensive care.

In this study, we aim to identify a transcriptomic signature of CT in both BD and MDD patients whose experiences during youth have been carefully assessed using the standardized childhood trauma questionnaire (CTQ). First, at the gene level, we conducted a differential expression analysis on: 1) RNA-seq data of lymphoblastoid immortalized cell lines generated from 37 BD patients and 20 healthy controls; and 2) RNA-seq data of peripheral blood monunuclear cells (PBMCs), generated from 30 MDD patients and 34 controls. Then, at a system level, we performed a weighted gene correlation network analysis (WGCNA), to detect specific or common modules and hub genes amongst both disorders. The expression of all genes within each module was summarized as eigengene to test association between module expression and CT.

Our preliminary results indicate interesting enrichments, with up-regulated genes involved in immune response in both BD and MDD, and down-regulated genes implicated in neuronal development among BD patients. Further characterization of genes whose expression is modified by the exposure to CT might provide a more comprehensive pathophysiological pathway leading from CT exposure to a higher severity in BD and MDD.

Keywords: Childhood Trauma, Bipolar Disorder, Major Depressive Disorder, WGCNA

1. References

[1] Oquendo, M. A., Currier, D., Liu, S.-M., Hasin, D. S., Grant, B. F., Blanco, C. (2010). Increased Risk for Suicidal Behavior in Comorbid Bipolar Disorder and Alcohol Use Disorders. The Journal of Clinical Psychiatry, 71(7), 902–909. https ://doi.org/10.4088/jcp.09m05198gry

[2] Aas, M., Henry, C., Andreassen, O. A., Bellivier, F., Melle, I., Etain, B. (2016). The role of childhood trauma in bipolar disorders. International Journal of Bipolar Disorders, 4(1). https ://doi.org/10.1186/s40345-015-0042-0

Iwakura Y., Fujisawa Y., Toyonaga N., Oral disintegrant tablet containing risperidone and method for producing the same, JP Patent, (2013) JP2013060392A.

[3] Etain, B., Lajnef, M., Brichant-Petitjean, C., Geoffroy, P. A., Henry, C., Gard, S.,Bellivier, F. (2016). Childhood trauma and mixed episodes are associated with poor response to lithium in bipolar disorders. Acta Psychiatrica Scandinavica, 135(4), 319–327. https ://doi.org/10.1111/acps.12684

Transcriptomic analysis of habenular asymmetries in the catshark *S. canicula*

Hélène MAYEUR[1], Maxence LANOIZELET[1], Ronan LAGADEC[1], Léo MICHEL[1], Christophe KLOPP[2], Bernard BILLOUD[3], Sylvie MAZAN[1]

[1] CNRS Sorbonne Universités, UMR 7232, Observatoire Océanologique, 66650 Banyuls/Mer, France
[2] Genotoul, INRA Toulouse, 31326 Castanet Tolosan cedex, France
[3] CNRS Sorbonne Universités, UMR 8227, Station Biologique, 29680 Roscoff, France

Corresponding author: sylvie.mazan@obs-banyuls.fr

Habenulae are bilateral epithalamic structures present in all vertebrates. They exhibit asymmetries, which extensively vary in nature and degree across vertebrates. Molecular analysis of habenular asymmetries has mostly been focused on zebrafish, which leaves the evolutionary mode underlying these variations poorly understood. In order to have a first estimate of the level of molecular asymmetries conservation in gnathostomes, we conducted a transcriptomic comparison between left and right differentiating habenulae in a cartilaginous fish exhibiting marked habenular asymmetries, the catshark *Scyliorhinus canicula* [1].

cDNA libraries were constructed from 2x3 RNA pools, each extracted from 15 manually dissected left and right habenulae at an advanced stage of differentiation. The libraries were sequenced using Illumina HiSeq to obtain a catshark habenula RNA-Seq dataset of 473 million reads. In the absence of a highly contiguous catshark genome, we used the DRAP pipeline [2] to generate a transcriptome de novo assembly including this and other publicly available catshark SRA datasets. The resulting contigs were then clustered with Corset [3] and SuperTranscripts [4] to decrease redundancy. This resulted in a reference database, termed SuperCatshark, containing 37,918 transcripts (N50=3,366 bp).

Left and right habenula reads were mapped on this reference transcriptome using the k-mer pseudo-mapping software package kallisto [5]. Paired statistical analyses of the pseudo-counts were performed using sleuth [6]. This produced a list of 682 putative differentially expressed contigs, of which 435 could be annotated by comparison with available databases (263 left-enriched, 172 right-enriched). Most of these genes are described for the first time as asymmetrically expressed in vertebrate habenulae. In order to validate these data, we carried in situ hybridizations on catshark habenula sections for about 50 transcripts selected among those exhibiting the highest q-values. For most of them, we observe highly asymmetric expression profiles, in line with the transcriptomic analysis results. The labeled territories also give insight into the sub-domain organization of the catshark habenulae, providing the first molecular map of habenulae in a chondrichthyan.

On-going work aims at (1) generating novel reference databases taking advantage of Unigene databases available in other chondrichthyans, in order to assess the impact of redundancy in the transcriptome used for read mapping, (2) comparing the molecular asymmetries characterized in the catshark with those reported in the zebrafish and (3) obtaining a genome-wide molecular map of the catshark habenulae using RNA tomography [7].

References

1. Ronan Lagadec, Laurent Laguerre, Arnaud Menuet, Anis Amara, Claire Rocancourt, Pierre Péricard, Benoît G. Godard, Maria Celina Rodicio, Isabel Rodriguez-Moldes, Hélène Mayeur, Quentin Rougemont, Sylvie Mazan and Agnès Boutet. The ancestral role of nodal signaling in breaking L/R symmetry in the vertebrate forebrain. *Nature Communications*. 6: 6686, 2015
2. Cédric Cabau, Frédéric Escudié, Anis Djari, Yann Guiguen, Julien Bobe, and Christophe Klopp. Compacting and correcting Trinity and Oases RNA-Seq de novo assemblies. *PeerJ*. 5:e2988, 2017.
3. Nadia M. Davidson and Alicia Oshlack. Corset: enabling differential gene expression analysis for de novo assembled transcriptomes. *Genome Biology*. 15:410, 2014.
4. Nadia M. Davidson, Anthony D. K. Hawkins and Alicia Oshlack. SuperTranscripts: a data driven reference for analysis and visualization of transcriptomes. *Genome Biology*. 18:148, 2017.
5. Nicolas L. Bray, Harold Pimentel, Páll Melsted and Lior Pachter. Near-optimal probabilistic RNA-seq quantification. *Nature Biotechnology*. 34:525–527, 2016.
6. Harold J. Pimentel, Nicolas Bray, Suzette Puente, Páll Melsted and Lior Pachter. Differential analysis of RNA-Seq incorporating quantification uncertainty. *Nature Methods*. 14:687–690, 2017.
7. F. Kruse, J. P. Junker, A. van Oudenaarden and J. Bakkers. Tomo-seq: a method to obtain genome-wide expression data with spatial resolution. Methods in Cell Biology. 135:299-307, 2016.

Transcriptomics Signature of Type I Narcolepsy (T1N)

Leila KHAJAVI[1,2,3], X-Hung NGUYEN[2], Raphael BERNARD-VALNET[2], Matthias ZYTNICKI[3], Roland LIBLAU[2]

[1] UPS Toulouse III: Ecole Doctorale: Biologie, Sante, Biotechnologies (BSB) 118 Rte de Narbonne, 31062 Toulouse, France
[2] INSERM U1043: Center for Pathophysiology Toulouse Purpan (CPTP), Université Toulouse III, CNRS UMR 5282, 31024 Toulouse, France
[3] INRA: Unité de Mathématiques et Informatique Appliquées de Toulouse (MIAT) Chemin de borde rouge, B.P. 52627 - 31326 CASTANET-TOLOSAN, Toulouse, France

Corresponding Author: roland.liblau@inserm.fr

Introduction

Narcolepsy (type 1) is a rare and severe sleep disorder characterized with the exclusive and extensive destruction of orexin-producing hypothalamic neurons [1]. The patients exhibit cataplexy (episodes of partial or complete paralysis of the voluntary muscles), sleep paralysis, hypnagogic hallucinations, as well as excessive daytime sleepiness and fragmented nocturnal sleep. Orexin-A and -B are small peptide neurotransmitters produced only by a cluster of neurons in the lateral hypothalamus which stimulate target neurons that promote wakefulness, regulate appetite and conserve energy. The mechanisms involved in the selective destruction of orexin neurons are not yet elucidated but its association with certain environmental triggers (vaccine against flu virus - Pandemrix), genetic susceptibility (strong association with the human leukocyte antigen – HLA), T-cell receptor (TCR) and other immune loci (such as CTSH, P2RY11, ZNF365, IFNAR1, TNFSF4) implicate an immunopathological process [2].

Experimental Design

To investigate whether an (auto)immune process could lead to narcolepsy development and to decipher the mechanisms involved in the selective loss of orexin neurons, we have developed a mouse model of immune-mediated narcolepsy (Orex-HA) to determine the molecular profiles of the hypothalamus infiltrating T-cells [3]. The Orex-HA mice are genetically engineered to express hemagglutinin (HA) of the H1N1 influenza virus as a self-antigen selectively on orexin neurons. In this experimental model, naive HA-specific CD4 and CD8 T-cells were injected into the mouse at day 0 and activated via immunization with the flu vaccine (Pandemrix) the following day. Vaccination results in the activation of HA-specific CD4 and CD8 T-cells which then migrate into the hypothalamus and target the orexin neurons expressing HA. At the peak of CNS infiltration (day 14), the mice were euthanized and the host (CD45.1-) and donor (CD45.1+) CD4 (CD44+CD62L-) and CD8 (CD44+CD62L-) T-cells were cell sorted (FACS) from the hypothalamus, spleen and cervical lymph nodes (cLNs). Total RNA was then isolated and sent for mRNA sequencing to better understand the transcriptomics signature of pathogenic T-cells in narcolepsy.

Results

Since the focus of this objective was to determine the molecular profiles of the hypothalamus-infiltrating T-cells, we combined the spleen and cLNs samples into “periphery” and limited the analysis to “Brain vs Periphery” for each of the T-cell populations. We have interrogated the differentially expressed genes derived from the comparison of the brain versus periphery for each of the four different cell types separately. These gene lists were evaluated via Ingenuity Pathway Analysis software (IPA) for canonical pathway analysis, KEGG pathway analysis (Webgestalt), and gene ontology (GO) analysis (Webgestalt). After reviewing the top 10 pathways, a concise gene list was selected based on immunological relevance and curiosity. This gene list will be prioritized for the first level of validation via QPCR.

References

[1] Liblau RS, Vassalli A, Seifinejad A, *et al.* Hypocretin (orexin) biology and the pathophysiology of narcolepsy with cataplexy. *Lancet Neurology*, 14:318-28, 2015.
[2] Miyagawa, T and Tokunaga, K. Genetics of narcolepsy. *Human Genome Variation*, 6:4, 2019.
[3] Bernard-Valnet R, Yshii L, Quériault C, *et al.* CD8 T cell-mediated killing of orexinergic neurons induces a narcolepsy-like phenotype in mice. *PNAS USA*, 113:10956-61, 2016.

UMI-VarCal: a new UMI-based variant caller that efficiently improves low-frequency variant detection in paired-end sequencing NGS libraries

Ahmad Abdel Sater [1,2,3], Pierre-Julien Viailly [2,3], Thierry Lecroq [1], Élise Prieur-Gaston [1], Élodie Bohers [2,3], Mathieu Viennot [2,3], Vinciane Marchand [2,3], Philippe Ruminy [2,3], Hélène Dauchel [1], Martine Becker [2,3], Hervé Tilly [2,3], Pierre Vera [1,2], Fabrice Jardin [2,3]

[1] Normandie Univ, UNIROUEN, LITIS EA 4108, 76000 Rouen, France

[2] Centre Henri Becquerel, 76000 Rouen, France

[3] Normandie Univ, UNIROUEN, INSERM U1245, Team "Genomics and Biomarkers of Lymphoma and Solid Tumors", 76000 Rouen, France

Corresponding Author: ahmad.abdel-sater@chb.unicancer.fr

Abstract

Due to recent advances in the field of oncology, and especially the increased use of liquid biopsy to monitor the tumor burden in the blood, the rise of new variant calling algorithms or strategies adapted to the low frequency variant detection has become a must. Because of PCR enrichment and sequencing technologies limitations, artifactual variants (sequencing and DNA polymerase errors) are also introduced at low frequencies making the distinction between real variants and artifactual ones a true challenge. However, the recent use of Unique Molecular Identifiers (UMI) in targeted sequencing protocols has offered a trustworthy approach to accurately call low frequency variants.

Here, we present UMI-VarCal, a new UMI-based variant caller with remarkably higher specificity compared to raw-reads-based variant callers. Although our variant caller is far from being the only one that uses UMI information to call variants, UMI-VarCal stands out from the crowd by not relying on Samtools to do its pileup. Instead, thanks to an innovative homemade pileup algorithm specifically designed to treat the UMI tags present in the reads, our variant caller surpasses the other variant callers (OutLyzer [1], DeepSNVMiner [2], MAGERI [3]) in terms of specificity. Furthermore, being developed with performance in mind, our tool is considerably more efficient than the other approaches in terms of execution time and memory consumption.

We illustrate the results obtained using UMI-VarCal through the sequencing of a cohort of patients suffering from lymphoma. Example of biopsy and plasma sequencing results will be discussed and UMI-VarCal sensitivity/specificity will be compared to other variant calling approaches.

References

[1] Muller E, Goardon N, Brault B, et al. OutLyzer: software for extracting low-allele-frequency tumor mutations from sequencing background noise in clinical practice. *Oncotarget*. 2016;7(48):79485–79493. doi:10.18632/oncotarget.13103

[2] Andrews TD, Jeelall Y, Talaulikar D, Goodnow CC, Field MA. DeepSNVMiner: a sequence analysis tool to detect emergent, rare mutations in subsets of cell populations. *PeerJ*. 2016;4:e2074. Published 2016 May 24. doi:10.7717/peerj.2074

[3] Shugay M, Zaretsky AR, Shagin DA, et al. MAGERI: Computational pipeline for molecular-barcoded targeted resequencing. *PLoS Comput Biol*. 2017;13(5):e1005480. Published 2017 May 5. doi:10.1371/journal.pcbi.1005480

Understanding Chemical-Genetic Interactions

Loan Vulliard, Michael Caldera and Jörg Menche
CeMM Research Center for Molecular Medicine of the Austrian Academy of Sciences,
Lazarettgasse 14, AKH BT 25.3, A-1090 Vienna, Austria

Corresponding author: lvulliard@cemm.oeaw.ac.at

1 Appreciation of perturbation combinations in biological systems

Biological systems, from cells to tissues and organisms, are highly complex and multi-scale. Studying their components in isolation cannot account for many properties that only emerge from their combination. To describe the cellular phenotype resulting from a multitude of molecular interactions, information about the different proteins and their relationships can be integrated in a bottom-up manner into an interactome network [1]. Unfortunately, we currently lack a systematic understanding of how independent perturbations influence each other, for example whether a drug treatment would improve a disease condition, or result in unexpected adverse reactions. The importance for a deeper understanding of context-dependent medications is emphasized by the fact that none of the top 10 best-selling drugs in the United States were effective in more than one out of four patients they were given to in 2015 [2]. We therefore aim for the construction of high-resolution directed interaction networks describing context-dependent drug response.

2 Morphological screen of combined perturbations

We are therefore performing and analyzing an arrayed morphological screen, combining genetic and chemical perturbations in a human epithelial cell line. This project involves, among others, some aspects from big biodata, image analysis, machine learning, systems biology and network science. Knockouts will be performed using the CRISPR-Cas9 system to simulate different genetic backgrounds with a high knockout efficiency and low off-target effect. We focus on 200 cellular perturbations that induce strong morphological responses and designed a single-guide RNA library targeting Rho-GTPases and actin cytoskeleton remodelers, as they play an essential role in cell morphology and are related to several rare monogenic diseases [3]. This insures the relevance of the approach, as our systematic precise knockout design is an efficient model to study these conditions. The 311 chemical compounds are chosen by compiling drugs and pharmacologically active small molecules affecting cell morphology through diverse mechanisms of action. For all the 62200 pairwise combinations of compounds and knockouts, microscopy images will be processed and analyzed in order to extract morphological features describing the cell.

3 Interpretation and generalization through the prism of the interactome

Using a vector-based approach, we can obtain a detailed landscape of the interactions between internal and external perturbations and infer the type and directionality of these interactions [4]. This will result in a perturbation interaction network that can then be explored and interpreted through the prism of the interactome of molecular interactions to identify rules that govern the superposition of intrinsic and extrinsic perturbations. Moreover, the interaction data can be integrated as a novel annotation layer of the global interactome that can be compared with prior knowledge. This approach harnesses the power of systems biology and network science to go beyond what can be concluded from the interactions individually, thus allowing for system-wide conclusions.

References

[1] Michael Caldera, Pisanu Buphamalai, Felix Müller, and Jörg Menche. Interactome-based approaches to human disease. *Current Opinion in Systems Biology*, 3:88–94, 6 2017.

[2] Nicholas J. Schork. Personalized medicine: Time for one-person trials. *Nature*, 520(7549):609–611, 4 2015.

[3] Tatyana Svitkina. The Actin Cytoskeleton and Actin-Based Motility. *Cold Spring Harbor perspectives in biology*, 10(1):a018267, 1 2018.

[4] Bernd Fischer, Thomas Sandmann, Thomas Horn, Maximilian Billmann, Varun Chaudhary, Wolfgang Huber, and Michael Boutros. A map of directional genetic interactions in a metazoan cell. *eLife*, 4, 3 2015.

Unraveling the rules of the Exon Junction Complex deposition with CLIP-seq

Toni PATERNINA[1], Auguste GENOVESIO[1], and Hervé LE HIR[1]

1 Institut de Biologie de l'ENS (IBENS), 46, rue d'Ulm, 75005, Paris, France
Corresponding Author: paternin@biologie.ens.fr

1 The Exon Junction Complex

The Exon Junction Complex (EJC), is a multi-protein complex deposited by the splicing machinery onto nascent mRNAs upstream of exon-exon junction [1][2]. The EJC has a central role in the post-transcriptional fate of mRNAs, including their splicing, localization, translation, and decay [1]. However, the comprehensive list of EJC targets, as well as the rules dictating its deposition onto mRNAs are still unknown.

2 CLIP-seq data analysis

Since the birth and development of Cross-linking and ImmunoPrecipitation coupled to deep sequencing (CLIP-seq) techniques, the study of RNA-binding protein targets has shifted from a few reporter genes to the scale of the transcriptome [3]. With these techniques, we obtain cDNA libraries of the protein-bound RNA fragments in a given cell type or tissue.

The standard CLIP-seq data analysis pipeline consists in peak calling, motif search, and accumulation of read signal relative to a known binding site (meta-analysis plots or RNA maps) [4]. However, we found that individual peaks display a low reproducibility rate, as shown by the Jaccard index of peaks from 84 CLIP-seq experiments (median = 0.20). To overcome this seemingly inherent low reproducibility of CLIP peaks, we propose a peak-calling independent approach to study EJC-specific signal quantitatively.

3 Results : quantitative comparison of EJC signal

The standard data analysis pipeline is essentially a qualitative study of protein-RNA interactions. In our approach, we count the number of reads within the region where EJC binding is expected (known as the canonical region) of all protein-coding exons. We thus obtain a *distribution* of the signal rather than the single curve obtained with an RNA map.

With thousands of observations, we are able to test differences in the distribution of EJC signal against negative controls, independently of peak detection. As an internal control, we compare it to a non-specific exon region known as non-canonical. As external controls, we use IP input control (non-specific RNA fragments) and RNA-seq data.

We observe that EJC signal is significantly stronger ($P < 0.001$, t-test) in the canonical region than in the non-canonical; it is slightly lower in input data ($P < 0.001$), and not different in RNA-seq data. This results validate our approach as a way to analyze quantitatively EJC-specific signal.

Furthermore, annotating exons according to different parameters (gene expression, exon and intron length, transcription rate, etc…) enables the analysis of EJC signal variation in relation to these features, leading to a *quantitative* model for the EJC deposition rules.

Our pipeline enables a global study of the EJC from a quantitative perspective, while bypassing the current low peak reproducibility.

4 References

[1] Le Hir, Saulière, Wang, The exon junction complex as a node of post-transcriptional networks. *Nat Rev Molecular Cell Biology.* 2016
[2] Hocq, Paternina, et al., Monitored eCLIP: high accuracy mapping of RNA-protein interactions. *Nucleic Acids Research.* 2018
[3] Lee, Ule, Advances in CLIP technologies for studies of protein-RNA interactions. *Mol Cell*. 2018
[4] Chakrabarti, et al., Data science issues in studying protein–RNA interactions with CLIP technologies. *Annu Rev Biomed Data Sci*. 2018

Unveiling the neoantigen landscape of malignant pleural mesothelioma using computational prediction and multiomics data

Arnaud Poret[1], Nicolas Alcala[1], Lise Mangiante[1], Francoise Galateau-Salle[2], Lynnette Fernandez Cuesta[1] and Matthieu Foll[1]
[1] International Agency for Research on Cancer, 150 cours Albert Thomas, 69008, Lyon, France
[2] Centre Léon Bérard, 28 Promenade Léa et Napoléon Bullukian, 69008, Lyon, France

Corresponding author: poreta@fellows.iarc.fr

Malignant Pleural Mesothelioma (MPM) is a deadly disease with most patients dying within 2 years after diagnosis. It is related to asbestos exposure, with a long latency between exposure and disease onset [1]. As the peak of asbestos use is yet to exceed the latency window, MPM incidence is expected to increase. MPM is a rare and understudied disease with limited therapeutic opportunities [2], although we [3] and others [4,5] have pointed to the potential benefit of immunotherapy.

As part of the French MESOMICS project [3], we have performed whole genome sequencing (WGS), RNA sequencing (RNAseq), and 850k methylation arrays (850K) on 112 MPM tumor samples and their matched normal. Analysis performed on these multiomics data provided information about copy number changes, rearrangements and somatic mutations, among others. It gave insights on the genetic variations seen in MPM tumors and their resulting neoantigen landscapes.

Given a tumor sample, its associated multiomics data were used with the computational tool Polysolver [6] for HLA typing. Once the HLA genotypes determined, the bioinformatics pipeline pVACseq [7] was used to predict putative neoantigens using MHCflurry [8], an algorithm for class I MHC binding affinity prediction. It resulted in a list of predicted neoantigens per typed HLA allele and tumor sample.

Combining HLA typing and subsequent neoantigen prediction would contribute to the endeavors deployed in immunotherapy for the treatment of malignant pleural mesothelioma [9]. Needless to mention that immunotherapy is, by definition, a medicine of precision. Adding the neoantigen layer would add the personalized dimension at the tumor level.

References

[1] Aude Lacourt, Emilie Leveque, Elie Guichard, Anabelle Gilg Soit Ilg, Marie-Pierre Sylvestre, and Karen Leffondre. Dose-time-response association between occupational asbestos exposure and pleural mesothelioma. *Occupational and Environmental Medicine*, 74(9):691–697, 2017.

[2] Anna C Bibby and Nick A Maskell. Current treatments and trials in malignant pleural mesothelioma. *The Clinical Respiratory Journal*, 12(7):2161–2169, 2018.

[3] Nicolas Alcala, Christophe Caux, Nicolas Girard, James D McKay, Francoise Galateau-Salle, Matthieu Foll, and Lynnette Fernandez-Cuesta. Redefining mesothelioma types as a continuum uncovers the immune and vascular systems as key players in the diagnosis and prognosis of this disease. *bioRxiv*, page 334326, 2018.

[4] Luana Calabro, Giulia Rossi, and Michele Maio. New horizons from immunotherapy in malignant pleural mesothelioma. *Journal of Thoracic Disease*, 10(Suppl 2):S322, 2018.

[5] Jordan Dozier, Hua Zheng, and Prasad S Adusumilli. Immunotherapy for malignant pleural mesothelioma: current status and future directions. *Translational Lung Cancer Research*, 6(3):315, 2017.

[6] Sachet A Shukla, Michael S Rooney, Mohini Rajasagi, Grace Tiao, Philip M Dixon, Michael S Lawrence, Jonathan Stevens, William J Lane, Jamie L Dellagatta, Scott Steelman, et al. Comprehensive analysis of cancer-associated somatic mutations in class i hla genes. *Nature Biotechnology*, 33(11):1152, 2015.

[7] Jasreet Hundal, Beatriz M Carreno, Allegra A Petti, Gerald P Linette, Obi L Griffith, Elaine R Mardis, and Malachi Griffith. pvac-seq: a genome-guided in silico approach to identifying tumor neoantigens. *Genome Medicine*, 8(1):11, 2016.

[8] Timothy J O'Donnell, Alex Rubinsteyn, Maria Bonsack, Angelika B Riemer, Uri Laserson, and Jeff Hammerbacher. Mhcflurry: open-source class i mhc binding affinity prediction. *Cell Systems*, 7(1):129–132, 2018.

[9] Linda Ye, Shaokang Ma, Bruce W Robinson, and Jenette Creaney. Immunotherapy strategies for mesothelioma-the role of tumor specific neoantigens in a new era of precision medicine. *Expert Review of Respiratory Medicine*, 13(2):181–192, 2019.

Using residues coevolution to search for protein homologs through alignment of Potts models

Hugo Talibart[1] and François Coste[1]
Univ Rennes, Inria, CNRS, IRISA, Campus de Beaulieu, 35042, Rennes, France

Corresponding author: hugo.talibart@irisa.fr

Thanks to sequencing technologies, the number of available protein sequences has considerably increased in the past years, but their functional and structural annotation remains a challenge. This task is classically performed *in silico* by retrieving well-annotated homologs with profile Hidden Markov Models (pHMMs), which are probabilistic models of families of homologous proteins capturing position-specific information with admissible insertion and deletion states. Two well-known software packages using pHMMs are widely used today : HMMER [1] aligns sequences to pHMMs to perform similarity searches, and HH-suite [2] takes it further by aligning pHMMs to pHMMs, enabling more sensitive remote homology searches.

Despite their solid performance, pHMMs are innerly limited by their positional nature. Yet, it is well-known that residues that are distant in the sequence can interact and co-evolve, e.g. due to their spatial proximity, resulting in correlated positions. Analyzing such correlations in a multiple sequence alignment by Direct Coupling Analysis [3], a statistical method to disentangle direct from indirect correlations, led to a breakthrough in the field of contact prediction [4]. Direct couplings are identified by inferring a Markov Random Field referred to as Potts model, and this model is of interest beyond its application in structure prediction. Indeed, its parameters can describe both positional conservation and direct couplings between residues of a protein. Such features drove us to examine Potts models for the purposes of modeling proteins and searching for their homologs.

In this poster, we focus on the use of Potts models for homology search, more specifically on our method for aligning and comparing Potts models. We present here our tool, named ComPotts, which formulates alignment of Potts models as an Integer Linear Programming problem and relies on a solver initially dedicated to pairwise protein alignment [5] to find efficiently the exact solution, and we present our first experimental results. Our ambition is to develop a package which would be equivalent to HH-suite but with Potts models rather than pHMMs, and to investigate on the added value of the direct couplings they provide.

Acknowledgements

HT is supported by a PhD grant from *Ministère de l'Enseignement Supérieur et de la Recherche* (MESR).
This work benefited from the support of the French government through the National Research Agency with regard to an investment expenditure program, IDEALG (ANR-10-BTBR-04).
We would like to warmly thank Inken Wohlers for providing us with her code.

References

[1] Sean R. Eddy. Profile hidden markov models. *Bioinformatics (Oxford, England)*, 14(9):755–763, 1998.

[2] Martin Steinegger, Markus Meier, Milot Mirdita, Harald Voehringer, Stephan J Haunsberger, and Johannes Soeding. Hh-suite3 for fast remote homology detection and deep protein annotation. *bioRxiv*, page 560029, 2019.

[3] Martin Weigt, Robert A White, Hendrik Szurmant, James A Hoch, and Terence Hwa. Identification of direct residue contacts in protein–protein interaction by message passing. *Proceedings of the National Academy of Sciences*, 106(1):67–72, 2009.

[4] Bohdan Monastyrskyy, Daniel D'Andrea, Krzysztof Fidelis, Anna Tramontano, and Andriy Kryshtafovych. New encouraging developments in contact prediction: Assessment of the casp 11 results. *Proteins: Structure, Function, and Bioinformatics*, 84:131–144, 2016.

[5] Inken Wohlers. *Exact Algorithms For Pairwise Protein Structure Alignment*. PhD thesis, Vrije Universiteit, 01 2012.

Vcf2Table: a VCF prettifier.

Pierre Lindenbaum, Matilde Karakachoff and
Richard Redon

L´ Institut du thorax, INSERM, CNRS, UNIV Nantes, Nantes, France.

May 13, 2019

The *Variant Call Format (VCF)* is the standard format for storing variants and genotypes. Reading all those details can be laborious, especially when there is a lot of genotypes or when there is a large number of items in the INFO column. To answer the need of easily reading the data in a terminal, we wrote a prettifier named *vcf2table*.

The tool displays the basic information about the variant (position,...), about each allele, the filters, the items in the INFO column. It displays information for the popular functional annotations like VEP or SNPeff : the format described in the VCF header is used to split the information in the INFO column and the information about each transcript is displayed in a table. The columns containing no information are removed. It displays the genotypes in a vertical table, making it easy to visualize a large number of samples. An option can be used to hide the 'hom-ref' or the 'no-call' genotypes to emphasize the variants (Figure 1).

The tool is available at http://lindenb.github.io/jvarkit/VcfToTable.html
.

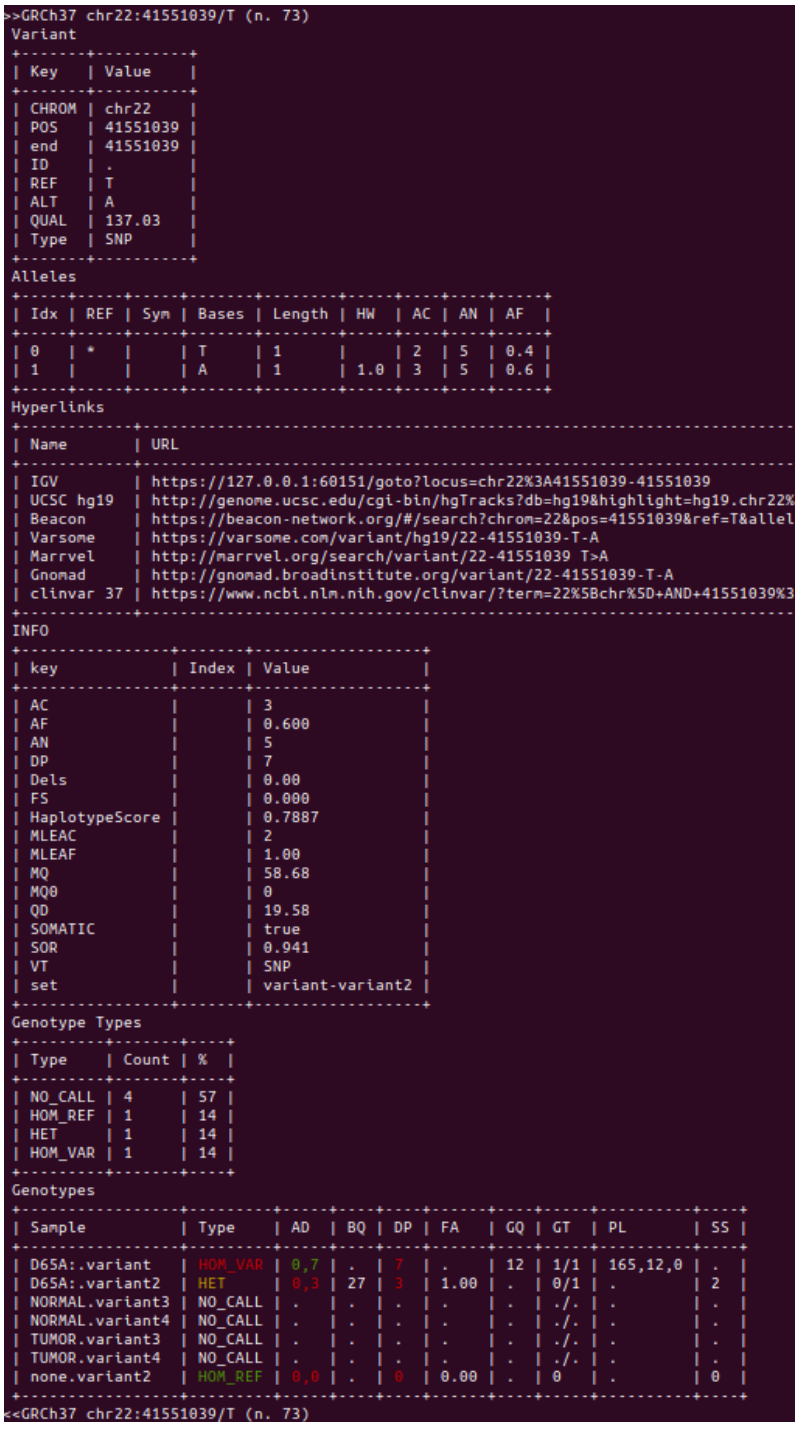


Figure 1: A screenshot of '*vcf2table*'.

Vidjil, une plateforme pour l'analyse des répertoires immunitaires

Florian Thonier[1], Mathieu Giraud[2] et Mikaël Salson[2]
[1] Consortium VidjilNet, Inria, 35000 Rennes, France
[2] CRIStAL (UMR 9189 CNRS, Université de Lille) 59 655 Villeneuve d'Ascq, France

Auteur référent: florian.thonier@inria.fr

1 Recombinaisons V(D)J et répertoire lymphocytaire

Les lymphocytes B et T jouent un rôle clé dans le système immunitaire *adaptatif*. Ces cellules peuvent être identifiées par une région de leur ADN appelé recombinaison V(D)J, propre à chaque lymphocyte, et qui permet la reconnaissance d'antigènes spécifiques [**?**]. Le mécanisme de recombinaison inclut une sélection de segments V, D et J parmi un pool spécifique pour chaque type et une juxtaposition de ces segments, avec des étapes aléatoires de délétions et d'insertions de nucléotides terminaux. De part leur spécificité, ces recombinaisons d'ADN peuvent servir de marqueurs d'évolution de maladies telles que les leucémies. Décrire quantitativement et qualitativement le répertoire lymphocytaire, permet ainsi d'améliorer le diagnostic et le suivi de certaines leucémies, et, plus généralement, d'aider à des avancées en hématologie et en immunologie [**?**].

2 La plateforme Vidjil

La plateforme open-source Vidjil (www.vidjil.org) analyse ces séquences d'ADN en identifiant et caractérisant ces recombinaisons V(D)J [**?**].

Vidjil-algo, analyse à haut-débit. Les reads sont rapidement regroupées en *clones* avec des méthodes sans alignement à base de graines espacées [**?**]. Ensuite, pour chaque clone, une analyse plus fine identifie les gènes V, D et J impliqués ainsi que les insertions, délétions et mutations spécifiques.

Plateforme web. L'utilisateur interagit avec une plateforme web composée d'un client (HTML/js/d3js) et d'un serveur (python/web2py) couplée à une base de données de patients, de runs de séquençage et d'expériences. Elle ou il peut ainsi lancer des analyses, visualiser les résultats, annoter ses observations et, dans un cadre clinique, identifier des marqueurs pronostics. La plateforme s'installe via des containers Docker déployés dans les hôpitaux, et un serveur public est accessible sur app.vidjil.org.

3 Routine hospitalière et consortium VidjilNet

Vidjil est utilisé à travers le monde par plus de 60 laboratoires, dont une trentaine réguliers. Depuis 2015, plus de 8 000 échantillons de diagnostic de leucémies aiguës lymphocytiques (LAL) ou de leucémies lymphocytaires chroniques (LLC) ont ainsi été analysés en routine hospitalière sur la plateforme. En 2019, la plupart des échantillons de diagnostic des LAL pédiatriques en France, Belgique, Italie et République tchèque sont ainsi analysés avec Vidjil.

Depuis 2018, le *consortium VidjilNet* (www.vidjil.net), hébergé par Inria, réunit développeurs et utilisateurs afin de pérenniser la maintenance et le support à la plateforme et de décider au mieux les évolutions à apporter au logiciel. Outre un focus sur le développement logiciel (2000 tests, intégration continue), le consortium s'efforce de développer et de maintenir la plateforme dans un cadre ouvert, éthique et réglementaire en vue d'une certification.

Nous remercions l'ensemble de nos utilisateurs en laboratoires cliniques ou de recherche.

Références

[1] Susumu Tonegawa, Somatic generation of antibody diversity. *Nature*, 302(5909), 575–581, 1983.

[2] Marc Duez et al., Vidjil : A web platform for analysis of high-throughput repertoire sequencing. *PLOS One*, 11(11), e0166126, 2016.

[3] Anton W. Langerak et al. High-throughput immunogenetics for clinical and research applications in immunohematology : Potential and challenges. *The Journal of Immunology*, 198(10), 3765–3774, 2017.

[4] Mathieu Giraud, Mikaël Salson et al., Fast multiclonal clusterization of V(D)J recombinations from high-throughput sequencing. *BMC Genomics*, 15(1), 409, 2014.

ViSEAGO: Easier data mining of biological functions organized into clusters using Gene Ontology and semantic similarity

Christelle HENNEQUET-ANTIER[1§], Aurélien BRIONNE[1§] and Amelie JUANCHICH[1§]

[1] BOA, INRA, Université de Tours, 37380, Nouzilly, FRANCE

[§] Contributed equally to this work

Corresponding Author: Christelle.hennequet-antier@inra.fr

Abstract

The main objective of *ViSEAGO* workflow is to carry out a data mining of biological functions and establish links between genes involved in the study. We developed *ViSEAGO* in R to facilitate functional Gene Ontology (GO) analysis of complex experimental design with multiple comparisons of interest. It allows to study large-scale datasets together and visualize GO profiles to capture biological knowledge. The acronym stands for three major concepts of the analysis: *Vi*sualization, *S*emantic similarity and *E*nrichment *A*nalysis of *G*ene *O*ntology. It provides access to the last current GO annotations, which are retrieved from one of NCBI EntrezGene, Ensembl or Uniprot databases for available species. *ViSEAGO* extends classical functional GO analysis to focus on functional coherence by aggregating closely related biological themes while studying multiple datasets at once. It provides both a synthetic and detailed view using interactive functionalities respecting the GO graph structure and ensuring functional coherence supplied by semantic similarity. *ViSEAGO* has been successfully applied on several datasets from different species with a variety of biological questions. Results can be easily shared between bioinformaticians and biologists, enhancing reporting capabilities while maintaining reproducibility.

ViSEAGO is publicly available on https://forgemia.inra.fr/umr-boa/viseago.

Visualizing metadata change in networks and / or clusters

Tanguy LALLEMAND[1], Sylvain GAILLARD[1], Sandra PELLETIER[1], Claudine LANDÈS[1], Sébastien AUBOURG[1]and Julie BOURBEILLON[1]

[1] IRHS, Agrocampus-Ouest, INRA, Université d'Angers, SFR 4207 QuaSaV, 49071, Beaucouzé, France

Corresponding Author: julie.bourbeillon@agrocampus-ouest.fr

1 Introduction

Among currents trends in various research fields, in particular in biology, is the increase in the scale at which studies are performed. This results from the wider spread of high-throughput experimental techniques such as transcriptomics or proteomics and the increase of the volume of publicly available datasets. For instance, the biology teams from the IRHS (Institut de Recherche en Horticulture et Semences) in Angers have been accumulating datasets of different natures (transcriptomic, biochemistry, physical measures, sensory analysis, etc.) regarding perennial, annual and biannual plants. However experiments are performed independently and resulting data are cross-analysed manually and a-posteriori by scientists [1]. Therefore the demand by biologists to integrate heterogeneous and large datasets from "omics" and phenotyping activities is rapidly increasing [2]. The classical approaches imply collating various data sets into a large data matrix and mine the matrix through methods such as building networks to represent relationships between items or clustering items into closely related groups. These are in turn visualized as graphs or heatmaps for instance.

2 Issue at hand

The first problem is the interpretation of the visualizations in biological terms. A standardized description of each dataset in the integrated matrix has to be available. These metadata are more and more well-defined through standard formats such as MIAME [3] and filled in with concepts from reference ontologies. However those have to be presented to the user in an interpretive way, which can be challenging when dealing with annotations extracted from large knowledge representations such as the GO [4]. Moreover building networks or clustering are generally iterative approaches. Kinetics are also regularly found among biological datasets. The presentation of the metadata is then not only a one shot operation but has also to take into account some kind of chronology through the data mining steps or the course of the biological process.

3 Contribution

In this context we are developing a web-based tool to present biologists with visualizations of ontology annotations associated with biological network graphs or cluster heatmaps through time including: (i) as a series of snapshots corresponding to successive steps and (ii) a representation of the difference between two steps. Our approach builds on methods such as GO enrichment analysis [5] and visualization of changes in networks [6].

References

[1] Mercedes Arguello Casteleiro et al. Deep learning meets ontologies: experiments to anchor the cardiovascular disease ontology in the biomedical literature. *Journal of Biomedical Semantics* (9:13). 2018.
[2] James Hendler. Data Integration for Heterogenous Datasets. *Big Data*. 2(4):205-215. 2014.
[3] Alvis Brazma et al. Minimum information about a microarray experiment (MIAME)—toward standards for microarray data. *Nature Genetics* (29):365–371. 2001.
[4] Michael Ashburner et al. Gene Ontology: tool for the unification of biology. *Nature Genetics* (25):25–29. 2000.
[5] Huaiyu Mi et al. Large-scale gene function analysis with the PANTHER classification system. *Nature Protocols* (8):1551–1566, 2013.
[6] Martin Rosvall and al. Mapping change in large networks. Plos One. 5(1): e8694 2010.

Which genome browser to use for my data ?

Franck BONARDI[1*], Loïc COUDERC[1*], Isabelle GUIGON[1*], Jean-Pascal MENEBOO[1*], Pierre PERICARD[1*], Hélène TOUZET[2]

[1] bilille, Université de Lille, Institut Pasteur de Lille, 59 655 Villeneuve d'Ascq cedex, France

[2] CNRS, CRIStAL, 59655 Villeneuve d'Ascq cedex, France

Corresponding Author: helene.touzet@univ-lille.fr

*All authors contributed equally.

The development of genome sequencing projects since the early 2000s has been accompanied by efforts from the scientific community to develop interactive graphical visualization tools, called *genome browsers* or *genome viewers*. This effort has been further intensified with the advent of high throughput sequencing and the need to visualize data as diverse as Whole Genome Sequencing, exomes, RNA-seq, ChIP-seq, variants, interactions, in connection with publicly available annotation information. Over the years, genome viewers have become increasingly sophisticated tools essential in data exploration and interpretation (1,2).

We have reviewed seven genome browsers, selected for their notoriety and their complementarity: Artemis (3), GIVE (4), IGB (5), IGV (6), Jbrowse (7), Tablet (8) and UCSC Genome Browser (9). For each of these tools, we have examined the following criteria: availability as a web server or desktop tool, easy installation, quality of documentation and learning difficulty, supported data types and file formats, processing of large files, modalities for sharing sessions, customization, interconnection with databases, metadata processing and display, quality of the navigation and exploration. This evaluation was performed on a wide variety of data types including genome sequences, read alignments, variant calling, features and quantitative tracks. Those datasets were selected from RNA-seq and ChIP-seq analysis on both model (human) and non-model (*Toxoplasma gondii*) organisms. Our analysis allows to distinguish common features shared by all tools, and to pinpoint specificities, strengths and weaknesses of each viewer.

The purpose of this work is to provide simple guidelines to help potential genome browser users to make the best choice for what they need. We also would like to take the opportunity of this presentation at JOBIM to compare our experience with that of other users.

References

1. Comparison of human (and other) genome browsers. T.S. Furey, *Hum Genomics*; 2(4): 266–270 (2006)
2. A brief introduction to web-based genome browsers. J.Wang et al., *Briefings in Bioinformatics*, 14(2):131–143 (2013)
3. Artemis: an integrated platform for visualization and analysis of high-throughput sequence-based experimental data. T. Carver, et al., *Bioinformatics*, 28;4;464-9 (2012)
4. GIVE: portable genome browsers for personal websites. X. Cao et al., *Genome Biology*, 19:92 (2018)
5. Integrated Genome Browser: Visual analytics platform for genomics. N. H. Freese et al., *Bioinformatics,* 32 (14):2089-95 (2016)
6. Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration. H. Thorvaldsdóttir et al., *Briefings in Bioinformatics* 14(2):178-92 (2013)
7. JBrowse: a dynamic web platform for genome visualization and analysis. R. Buels et al., *Genome Biology*, 17:66 (2016)
8. Using Tablet for visual exploration of second-generation sequencing data. I. Milne et al, *Briefings in Bioinformatics*, 14-2:193-202 (2013)
9. The UCSC Genome Browser database: 2019 update. M. Haeussler et al., *Nucleic Acids Research*, 47-D1 : D853–D858 (2019)

Évaluation de la qualité et comparaison des assemblages des génomes

Anna Tran[1,2], Sèverine Bérard[1] et Anne-Muriel Arigon Chifolleau[2]
[1] ISEM, Université de Montpellier, CNRS, IRD, EPHE, place E. Bataillon, 34095, Montpellier, France
[2] LIRMM, UMR 5506 - CNRS, Université de Montpellier, 161 rue Ada, 34095, Montpellier, France

Auteur référent: anna.tran@etu.umontpellier.fr

Abstract

Depuis le progrès des technologies de séquençage, la reconstruction de génomes reste un des problèmes majeurs en bioinformatique. Elle est composée de différentes étapes qui aboutissent dans le meilleur des cas à un génome complètement assemblé (Figure 1).

Fig. 1. Pipeline de reconstruction d'un génome

Il est possible de retrouver des génomes assemblés dans différentes bases de données dont Assembly créée par NCBI et ENA par EMBL. Celles-ci permettent d'obtenir des informations relatives aux assemblages notamment le niveau d'assemblage atteint ("génome complet", "chromosome", "scaffolds" ou "contigs"). La plupart des assemblages s'arrêtent au niveau "scaffolds" puisqu'il n'est pas toujours possible d'effectuer l'étape de finition. Une des difficultés est que les scaffolds sont particulièrement sujets aux erreurs car générés à partir de reads courts[1]. De plus, les données actuelles contiennent des erreurs dans les reads (substitution, insertion, ...), des zones difficiles à séquencer et des "données brutes" différentes. Par ailleurs, plus d'une équipe peut travailler sur un même génome. Il est donc possible d'avoir plusieurs jeux de données différents provenant de multiples sources. Ainsi, pour chaque organisme, plusieurs assemblages d'un génome sont susceptibles d'être disponibles. Il est alors important de pouvoir évaluer de manière correcte la qualité des assemblages afin de sélectionner les plus pertinents mais aussi de s'appuyer sur ces différents assemblages. L'Assemblathon 2[2] a suggéré que faire du "méta-assemblage" en combinant plusieurs assemblages d'un même génome pourrait permettre d'en améliorer la qualité. En effet, actuellement il est peu probable d'obtenir un assemblage complet qui ne contienne aucune erreur [3].

L'objectif de ces travaux sera d'évaluer et de comparer différents assemblages d'un même génome. Pour cela, dans un premier temps, nous allons développer des méthodes permettant de comparer les différents assemblages pour obtenir des distances soit en utilisant leurs données brutes (séquences) soit en prenant en compte une sélection de critères de qualité. Et dans un second temps, une interface de visualisation permettra de mettre en avant les points forts et les points faibles de chaque assemblage en combinant les résultats des méthodes développées. Le travail effectué s'insère dans le cadre d'un projet qui, à terme, devrait mener au développement d'une application de "méta-assemblage" proposant une nouvelle stratégie en se basant sur la qualité des assemblages.

Références

[1] Martin Hunt, Chris Newbold, Matthew Berriman, and Thomas D. Otto. A comprehensive evaluation of assembly scaffolding tools. *Genome Biology*, 15(3) :R42, Mar 2014.

[2] Keith R Bradnam, Joseph N Fass, Anton Alexandrov, Paul Baranay, Michael Bechner, Inanç Birol, Sébastien Boisvert, Jarrod A Chapman, Guillaume Chapuis, Rayan Chikhi, et al. Assemblathon 2 : evaluating de novo methods of genome assembly in three vertebrate species. *GigaScience*, 2(1) :10, 2013.

[3] Hind Alhakami, Hamid Mirebrahim, and Stefano Lonardi. A comparative evaluation of genome assembly reconciliation tools. *Genome biology*, 18(1) :93, 2017.

Index

> Nantes
2—5
juillet

JOBIM 2019

JOURNÉES OUVERTES DE BIOLOGIE INFORMATIQUE & MATHÉMATIQUES

Thèmatiques :

Biologie structurale
Biologie des systèmes
Epidémiologie Génétique
Evolution/Phylogénie
Génomique/Métagénomique
Sciences des données

Keynotes :

Chloé-Agathe Azencott, Paris
Alexander Bockmayr, Berlin
Alessandra Carbone, Paris
Olivier Delaneau, Lausanne
Christophe Dessimoz, Lausanne
Juliette Martin, Lyon

https://jobim2019.sciencesconf.org