KaruBioNet: a network and discussion group for a better collaboration and structuring of bioinformatics in Guadeloupe (French West Indies)

David Couvin, Alexis Dereeper, Damien Meyer, Christophe Noroy, Stanie Gaete, Bernard Bhakkan, Nausicaa Poullet, Sarra Gaspard, Etienne Bezault, Isabel Marcelino, et al.

Bioinformatics Advances, 2022, 1–7
https://doi.org/10.1093/bioadv/vbac010
Advance Access Publication Date: 18 February 2022
Feature

Community report

KaruBioNet: a network and discussion group for a better collaboration and structuring of bioinformatics in Guadeloupe (French West Indies)

David Couvin [1,*,†], Alexis Dereeper[1,†], Damien F. Meyer [2,3], Christophe Noroy[4], Stanie Gaete[5], Bernard Bhakkan[6], Nausicaa Poullet[7], Sarra Gaspard[8], Etienne Bezault [9], Isabel Marcelino[1], Ludovic Pruneau[10], Wilfried Segretier[11], Erick Stattner[11], Damien Cazenave[1], Maëlle Garnier[1], Matthieu Pot[1], Benoît Tressières[12], Jacqueline Deloumeaux[5,6], Sébastien Breurec[1,12,13], Séverine Ferdinand[1], Silvina Gonzalez-Rizzo[10] and Yann Reynaud[1]; for the KaruBioNet Team[‡]

[1]Unité Transmission, Réservoir et Diversité des Pathogènes, Institut Pasteur de Guadeloupe, Les Abymes, Guadeloupe 97139, France, [2]CIRAD, UMR ASTRE, Petit-Bourg, Guadeloupe 97170, France, [3]ASTRE, Univ Montpellier, CIRAD, INRAE, Montpellier 34000, France, [4]Développement, Analyse, Transfert et Application (DATA), Lamentin, Guadeloupe 97129, France, [5]Karubiotec Centre de Ressources Biologiques-UF 0216, CHU de la Guadeloupe, Pointe-à-Pitre 97110, France, [6]Registre des cancers de Guadeloupe, CHU de la Guadeloupe, Pointe-à-Pitre 97110, France, [7]URZ Recherches Zootechniques, INRAE, Petit-Bourg, Guadeloupe 97170, France, [8]Laboratoire COVACHIMM2E EA3592, Université des Antilles, Pointe-à-Pitre, Guadeloupe 97110, France, [9]UMR BOREA (MNHN, CNRS-7208, IRD-207, Sorbonne Université, UCN, UA), Université des Antilles, Pointe-à-Pitre, Guadeloupe 97110, France, [10]Équipe « Biologie de la mangrove » UMR7205 « ISYEB » MNHN-CNRS-Sorbonne Université-EPHE-UA, UFR SEN Département de Biologie, Université des Antilles, Pointe-à-Pitre, Guadeloupe 97110, France, [11]Laboratoire de Mathématiques Informatique et Applications (LAMIA), Université des Antilles, Pointe-à-Pitre, Guadeloupe 97110, France, [12]Centre d'Investigation Clinique Antilles Guyane, Inserm CIC 1424, Les Abymes, Pointe-à-Pitre, Guadeloupe 97110, France and [13]Faculté de Médecine Hyacinthe Bastaraud, Université des Antilles, Pointe-à-Pitre, Guadeloupe 97110, France

*To whom correspondence should be addressed.
[†]The authors wish it to be known that, in their opinion, the first two authors should be regarded as Joint First Authors.
[‡]A list of other members of the KaruBioNet team is provided in the Acknowledgements.
Associate Editor: Nicola Mulder

Received on September 27, 2021; revised on January 24, 2022; editorial decision on January 27, 2022; accepted on February 9, 2022

Abstract

Summary: Sequencing and other biological data are now more frequently available and at a lower price. Mutual tools and strategies are needed to analyze the huge amount of heterogeneous data generated by several research teams and devices. Bioinformatics represents a growing field in the scientific community globally. This multidisciplinary field provides a great amount of tools and methods that can be used to conduct scientific studies in a more strategic way. Coordinated actions and collaborations are needed to find more innovative and accurate methods for a better understanding of real-life data. A wide variety of organizations are contributing to KaruBioNet in Guadeloupe (French West Indies), a Caribbean archipelago. The purpose of this group is to foster collaboration and mutual aid among people from different disciplines using a 'one health' approach, for a better comprehension and surveillance of humans, plants or animals' health and diseases. The KaruBioNet network particularly aims to help researchers in their studies related to 'omics' data, but also more general aspects concerning biological data analysis. This transdisciplinary network is a platform for discussion, sharing, training and support between scientists interested in bioinformatics and related fields. Starting from a little archipelago in the Caribbean, we envision to facilitate exchange between other Caribbean partners in the future, knowing that the Caribbean is a region with non-negligible biodiversity which should be preserved and protected. Joining forces with other Caribbean countries or territories would strengthen scientific collaborative impact in the

region. Information related to this network can be found at: http://www.pasteur-guadeloupe.fr/karubionet.html. Furthermore, a dedicated 'Galaxy KaruBioNet' platform is available at: http://calamar.univ-ag.fr/c3i/galaxy_karubionet.html.

Availability and implementation Information about KaruBioNet is availabe at: http://www.pasteur-guadeloupe.fr/karubionet.html

Contact: dcouvin@pasteur-guadeloupe.fr

Supplementary information: Supplementary data are available at *Bioinformatics Advances* online.

1 Introduction

Rapid advances in sequencing technologies over the past quarter-century have led to substantial reductions in the cost of genome sequencing, producing huge amounts of heterogeneous data. Sequencing and biological associated data become more affordable and available. However, the information generated is difficult to analyze. Then, new tools and methods are therefore needed.

As a French overseas territory, Guadeloupe is part of the European Union (EU) and belongs to its outermost regions. This archipelago is a small territory with an important local presence of national research organizations dealing with sequencing data. Bioinformatics is a multidisciplinary field that provides a large number of tools and methods that can be used to conduct scientific studies in a more strategic way. Until now, in Guadeloupe, this recent discipline was structured and coordinated at laboratory scale. In fact, most scientists used to favor the help of colleagues located outside the region and often long associated with the parent organization (mostly in mainland France) to perform the sequencing and analysis of their data. Although we can approve this practice for various needs of collaborations and strengthening of different projects, we believe that it would be beneficial for actors in the region to be able to access various resources and support platforms in bioinformatics and data sharing. Resources and networking are of primary interest to better establish collaboration between scientists (Hazbón *et al.*, 2018).

In the Caribbean, a bioinformatics and biostatistics network coordinating actions and collaborations was needed to find more innovative and accurate methods. As Karukera ('the island of beautiful waters') is the Native American name of Guadeloupe, we decided to create in January 2019 the Karukera Bioinformatics Network: KaruBioNet. However, we cannot omit lessons learnt from other renowned networks before establishing KaruBioNet such as EMBnet and H3ABioNet (D'Elia *et al.*, 2009; Mulder *et al.*, 2016).

Several organizations contribute to KaruBioNet. This network includes scientists from the Institut Pasteur de la Guadeloupe, the Université des Antilles (UA), the Centre de coopération internationale en recherche agronomique pour le développement (CIRAD), the Institut national de recherche pour l'agriculture, l'alimentation et l'environnement (INRAE), the Centre d'Investigation Clinique Antilles Guyane (CIC), the Institut national de la santé et de la recherche médicale (Inserm), the KarubiotecTM (Biological resources Center of University Hospital of Guadeloupe) and the Région Guadeloupe. The purpose of this group is to promote collaboration and mutual aid between people from different disciplines (biology, computer science, health informatics, biostatistics, mathematics, chemistry, epidemiology, ecology, clinical research, etc…). This transdisciplinary network is a platform for discussion, sharing, training and support between researchers, engineers and students interested in bioinformatics and related fields. Here, we aimed to establish and highlight the contribution of KaruBioNet to build solidarity links that allow better collaboration between researchers from various fields in Guadeloupe. Information related to this network can be found at: http://www.pasteur-guadeloupe.fr/karubionet.html. Furthermore, a dedicated 'Galaxy KaruBioNet' platform allowing to facilitate bioinformatic analyses is available at: http://calamar.univ-ag.fr/c3i/galaxy_karubionet.html. We also benefit from the 'Exocet' High-Performance Computing (HPC) facility of the UA to perform computing calculations.

2 Themes, tools and developments of a transdisciplinary network

2.1 Major common themes shared by laboratories

The main purpose of this group is to bring together people who share common problems to better collaborate together. The KaruBioNet network aims to improve the development of bioinformatics at the local or regional level for a better understanding and analysis of real-life data. The major common themes shared by laboratories belonging to the network are microbial evolutionary history, antimicrobial resistance, virulence mechanisms, systems biology, genotyping and data science (Fig. 1). The KaruBioNet network particularly aims to help researchers in their studies related to metagenomics, proteomics, genomics (or other 'omics'), as well as more general aspects relating to data analysis, integration and interpretation. In order to structure discussions and exchanges between researchers, we have implemented different topics (which could evolve in the future): (i) omics sciences; (ii) artificial intelligence and machine learning; (iii) biochemical analyses; (iv) geographic information systems; (v) databases and software development; (vi) biostatistics; and (vii) epidemiology. A video channel was created to disseminate training materials in French as well as presentations or demonstrations related to bioinformatics (details are provided in Supplementary Data).

2.2 Needs and challenges specific to omics sciences

Concerted actions are needed to strengthen and maintain computational and sequencing resources between institutions. We also intend to promote the development of dedicated bioinformatics pipelines and tools specific to omics data analysis. Several future challenges are foreseeable: (i) efforts will be necessary to find sustainable financial support to maintain the network; (ii) new and innovative strategies will be needed to share data and analytical pipelines (like Galaxy); and (iii) improved training materials will promote the development of bioinformatics in our region and elsewhere.

This would allow analytical strategies to be built around the One Health concept defined as a collaborative, multisectoral and transdisciplinary approach working at local, regional, national and global levels, to achieve optimal health outcomes and recognize the interconnection between people, animals, plants and their common environment (Lerner and Berg, 2017). This approach would make the challenge relevant for a better understanding and surveillance of human, plant or animal health and diseases (Gruel *et al.*, 2021).

3 Sequencing techniques

Sequencing devices are currently available in Guadeloupe. Devices such as Nanopore MinION (https://nanoporetech.com/products/minion) or Illumina MiSeq (https://emea.illumina.com/systems/sequencing-platforms/miseq.html) can be used for real-time sequencing projects.

Various sequencing projects are ongoing in several labs involved in the KaruBioNet. The impact of bacterial genomes such as *Klebsiella pneumoniae*, *Escherichia coli* or *Enterobacter cloacae* are being studied in depth in our environment (notably their ability to be resistant to some antibiotics; Pot *et al.*, 2021). Moreover, to study bacterial adaptations to environmental changes or the interaction with organisms in the same environment, RNA sequencing is a powerful tool. It makes it possible to study the differential expression of messenger RNAs and to identify potential regulators of non-coding RNAs. RNA sequencing

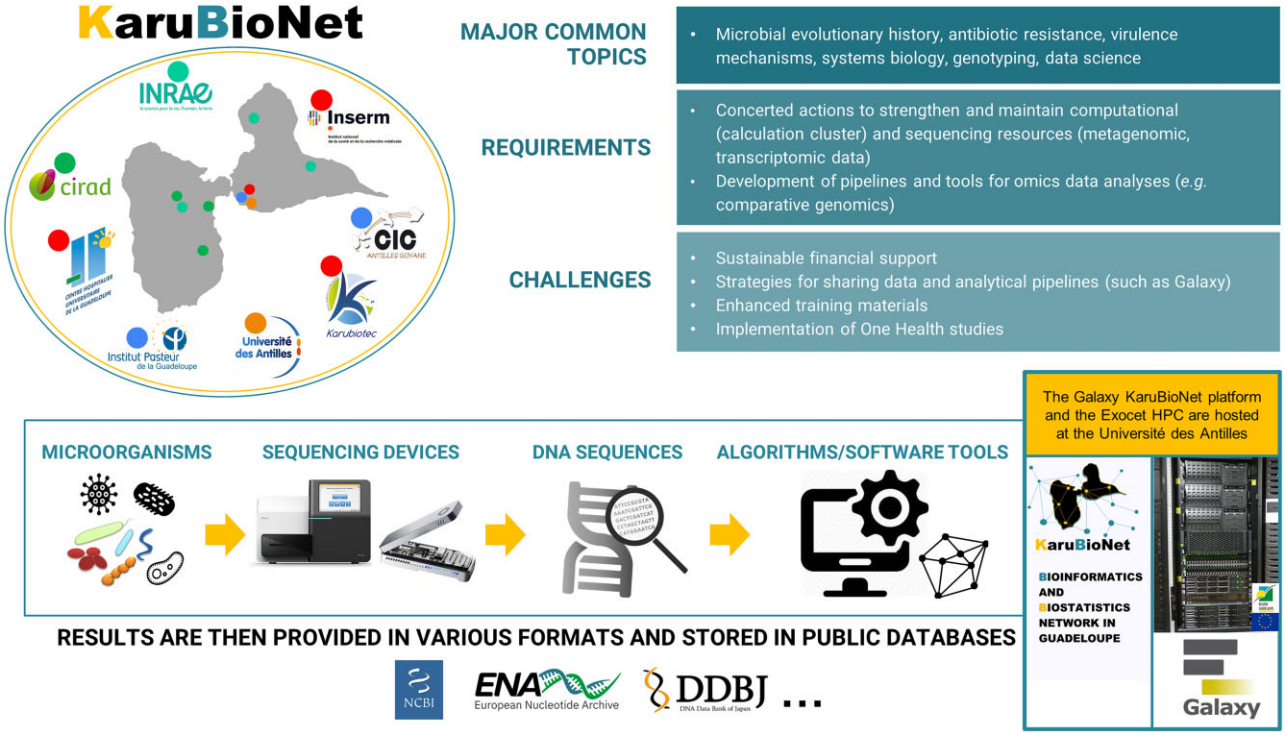

Fig. 1. Mapping showing major common themes shared by research institutions involved in the KaruBioNet. The colored dots represent the location of each institution on the map. This figure also shows a simplified workflow for sequencing data analysis. Resources and platforms are mainly located at the UA

generates a lot of data and requires the help of specific software accessible in the KaruBioNet network. A simplified workflow for sequencing data analysis is shown in Figure 1. A wide variety of heterogeneous sequencing data is produced and analyzed within the network (details are provided in Supplementary Data).

4 Exocet computing cluster

The 'Exocet' HPC facility of the UA provides us a great capacity for analyzing whole-genome sequencing (WGS) data. With almost 1000 CPU cores and 7349 GB RAM, Exocet cluster contains powerful nodes and NVidia graphics cards allowing it to perform bioinformatics analyses for all members of the KaruBioNet team.

Resources are shared among all members of KaruBioNet via access to Exocet HPC. Users must make a request to the Center Commun de Calcul Intensif of the UA (http://calamar.univ-ag.fr/c3i/exocet.html). All members of the network can access Exocet and other services free of charge. KaruBioNet administrators manage dedicated servers and other services. We promote our various services where possible to facilitate their use.

Exocet contains:

- Two front-end servers named 'exocet1' and 'exocet2'. When we log into Exocet, we are actually logging into one of the two servers.
- Twenty-five 'Calculation' nodes (node01 to node25) are used for general calculations on Intel processors. Each node contains two Intel processors of 18 cores each, making a total of 36 cores for one node, and 192 GB of RAM.
- A 'large RAM' node (mem01) with the same two Intel processors as a 'Calculation' node but with 1536 GB of RAM memory.
- A 'V100' node (gpu01) having the same characteristics as a 'Calculation' node but with in addition two Nvidia TESLA V100 graphics cards.
- A 'T4' node (gpu02) similar to the 'V100' but with two NVidia T4 graphics cards.
- Dual-processor servers each with 256 GB of RAM memory and a QUADRO P5000 graphics card.

The mobaXterm application (https://mobaxterm.mobatek.net/) can be used to get connected to the computing cluster.

5 Galaxy KaruBioNet platform

We have started the installation of a local Galaxy platform (http://galaxyproject.org/; Afgan *et al.*, 2018) that will allow researchers and students who are not comfortable with command-line interfaces to perform bioinformatic analyzes in a user-friendly manner. This local platform is available online from a calculation server of the UA. Users need to register to the Galaxy platform using a login (email address) and a password (http://calamar.univ-ag.fr/c3i/galaxy_karubionet.html). This local Galaxy instance (hosted at the UA) can be considered as a showcase reflecting different themes and activities of the network (Fig. 1).

5.1 Dedicated specialized workflows

5.1.1 Bacterial pangenomics

A ready-to-use workflow is available for Galaxy KaruBioNet users to process a complete analysis of the bacterial pangenome/core-genome. Starting from a collection of strains for which WGS reads are available, the user will be able to process successively analytical steps from raw reads to fully assembled and annotated genomes, and then a comparative genomics analysis to finally obtain both a pangenome matrix image and a Circos (Krzywinski *et al.* 2009) representation of pangenome ready for publication (Fig. 2). The first part of the workflow takes advantage of the Galaxy dataset collections to reconstruct and annotate complete individual genomes for each strain (each program will be executed as many times as there are strains to analyze):

1. *De novo* assembly of Illumina reads with Unicycler (Wick *et al.*, 2017).

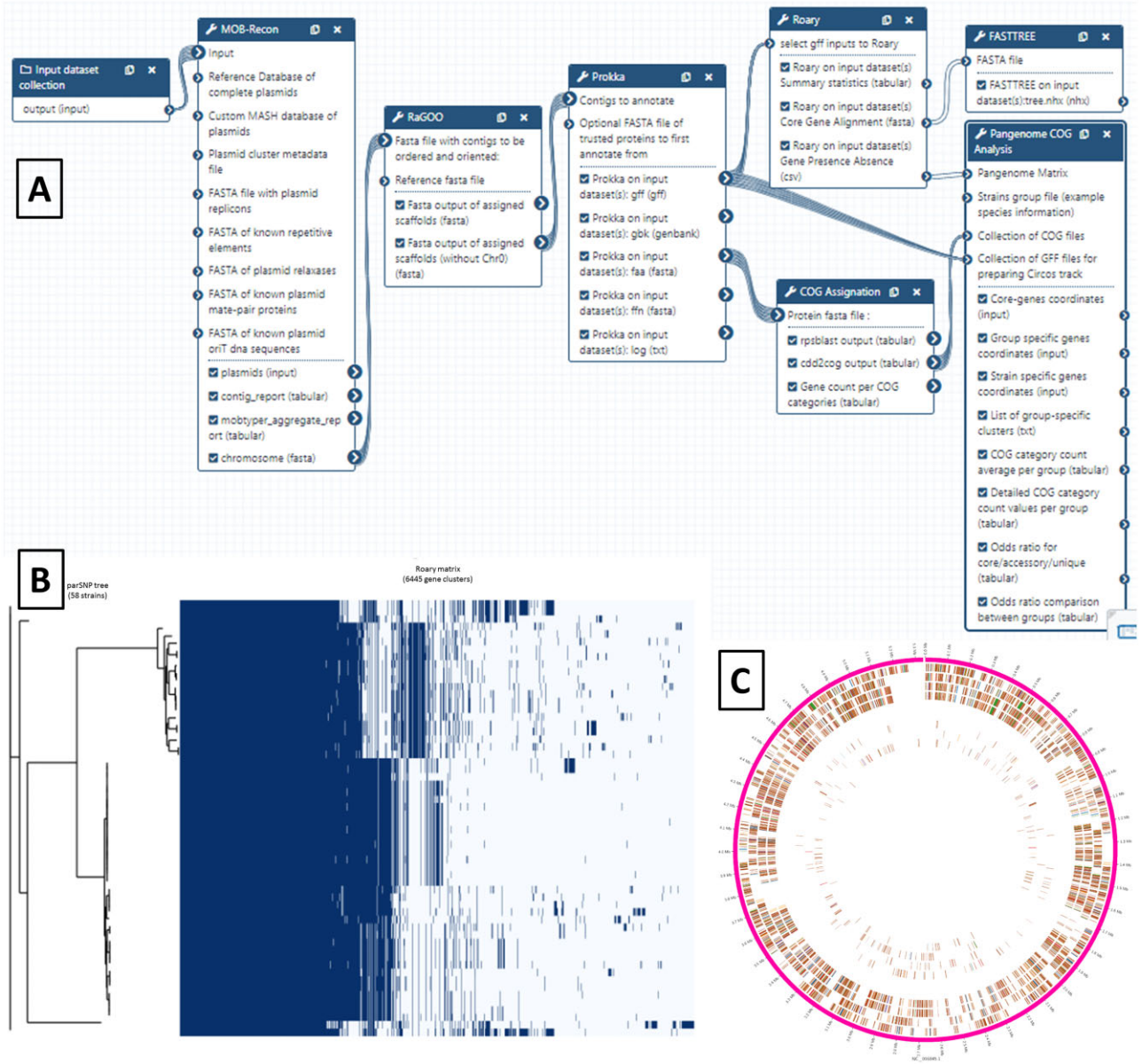

Fig. 2. Galaxy workflow and its graphical outputs for pangenome analysis of bacterial strains. (**A**) Bacterial pangenomics Galaxy workflow. (**B**) Gene presence/absence matrix facing the phylogenetic tree based on core-genome alignment. (**C**) Circos representation of core genes and specific genes

2. Assignation of contigs to plasmids or chromosomal with MOB-Recon (Robertson and Nash, 2018).
3. Anchoring and ordering of scaffolds using RaGOO (Alonge *et al.*, 2019), in order to establish a complete pseudomolecule for the chromosomal genome. This step requires an external reference genome.
4. Structural and functional annotation of reconstructed genomes using Prokka (Seemann, 2014).
5. COG (Cluster of Orthologs Groups; Galperin *et al.*, 2021) assignation of genes.

The second part of the workflow process is linear (each program will be executed only once) and allows the comparison of genomes:

1. Pangenome analysis using Roary (Page *et al.*, 2015).
2. Pangenome matrix representation using roary_plots.
3. Core-genome-based phylogeny using FastTree (Price *et al.*, 2010).
4. Circular representation of core genes and strain-specific genes positioned on a reference genome, using Circos.

5.1.2 *Mycobacterium tuberculosis* genomic analyses

Tuberculosis (TB) is an infectious disease caused by bacteria belonging to the *Mycobacterium tuberculosis* complex (MTBC). Interestingly, several Galaxy tools exist to study MTBC strains from available sequence reads or assembled genomes. Resources related to *M.tuberculosis* were tentatively listed in the MTBCtools list (Couvin *et al.*, 2021) in which tools were classified into several thematic sections/categories such as biological databases, drug resistance or prediction/classification tools. The Galaxy ToolShed (https://toolshed.g2.bx.psu.edu/) contains several TB tools such as Galru (Page *et al.*, 2020), lorikeet spoligotyping (Cohen *et al.*, 2015), Mykrobe (Hunt *et al.*, 2019), SpoTyping (Xia *et al.*, 2016) and TB-Profiler (Phelan *et al.*, 2019). We have recently added MIRUReader (Tang and Ong, 2020) to the Galaxy ToolShed. These tools could be used to produce genotyping information such as

spoligotypes (https://fr.wiktionary.org/wiki/spoligotype) or Mycobacterial interspersed repetitive units-variable number of tandem DNA repeats (MIRU-VNTRs), and detect drug resistance profiles from MTBC DNA sequences.

An example of a workflow to study the genotyping of *M.tuberculosis* strains and their association with antibiotic resistance would be as follows:

(i) users can analyze their sequence reads using the lorikeet spoligotyping program to obtain spoligotypes matching against their reads. Furthermore, they can use TB-Profiler program to determine the drug resistance profile from reads; (ii) they can then use a *de novo* assembly program such as Shovill (https://github.com/tseemann/shovill), SPAdes (Bankevich *et al.*, 2012) or Unicycler; and (iii) MIRUReader tool can finally be used to determine 24-loci MIRU-VNTRs profiles from preassembled data.

5.2 Other examples of modules/workflows

5.2.1 Simple phylogeny workflow

Phylogenetic analyzes are widely used to study biological data and other scientific data. From a Multi-FASTA file, users can perform a multiple alignment using a multiple alignment program such as MAFFT (Katoh and Standley, 2013). Users can then perform a phylogenetic analysis with a dedicated program such as FastTree, PhyML (Guindon *et al.*, 2010) or RAxML (Stamatakis, 2014). Note that user-friendly approaches such as Phylogeny.fr or NGPhylogeny.fr (Dereeper *et al.*, 2008; Lemoine *et al.*, 2019) already exist to perform phylogenetic analyses. The phylogenetic trees obtained can be annotated using iTOL (Letunic and Bork, 2021).

5.2.2 Construction of a variant call format file and single nucleotide polymorphism analyses

To construct a variant call format file, users can choose the BWA-MEM (Li, 2014) tool to map their sequence reads against a reference genome, then they can use Samtools/mpileup to perform a multi-way pileup of variants (Li *et al.*, 2009). Finally, they can use the VarScan (Koboldt *et al.*, 2012) program for variant detection. Other Galaxy pipelines such as Snippy (https://github.com/tseemann/snippy) could be used for single nucleotide polymorphism analyses.

5.2.3 Automated pipeline for the search for resistance, plasmid and virulence genes

We have developed a pipeline called catchSequenceInfo that allows us to get resistance, virulence and plasmids, as well as multilocus sequence typing (MLST; Jolley *et al.*, 2018) information from DNA sequences. This tool uses: (i) ABRicate (https://github.com/tseemann/abricate) with ResFinder (Zankari *et al.*, 2012), PlasmidFinder (Carattoli *et al.*, 2014) and VFDB (Liu *et al.*, 2019) databases to predict resistance, plasmid and virulence genes; as well as (ii) MLST (https://github.com/tseemann/mlst) to get allele IDs and MLST number. catchSequenceInfo is freely available through the Galaxy KaruBioNet instance and it has been placed in the Galaxy ToolShed. Note that the pMLST tool (Carattoli *et al.*, 2014) has also been installed in our Galaxy instance as well as in the ToolShed.

5.3 A local galaxy training focused on metagenomics

Metagenomic analyses supported by high-throughput sequencing provide a method to evaluate the microbial community in terms of both taxonomy and potential functioning.

A local Galaxy training took place on June 17, 2021 at the UA. During this training session, we used the Galaxy Training webpage (https://training.galaxyproject.org/) to practice with a tutorial titled Galaxy 101 (https://training.galaxyproject.org/training-material/topics/introduction/tutorials/galaxy-intro-101/tutorial.html). Then, we used another tutorial to perform metabarcoding analyses using the FROGS pipeline (Escudié *et al.*, 2018). This pipeline has been shown to be very effective for metagenomics studies. It could also be coupled with the analysis pipeline used in SHAMAN (Volant *et al.*, 2020).

6 Databases and data warehouses

Some developments are ongoing to construct dedicated databases or data warehouses.

The SITVIT databases and other resources (Couvin *et al.*, 2017, 2022, 2019, 2020; Demay *et al.*, 2012) developed by researchers at the Institut Pasteur de la Guadeloupe have provided a more complete global view of the molecular epidemiology of MTBC for public health surveillance. These resources are available online and may be of use to researchers working specifically on these bacterial organisms.

Bacterial pathogens have evolved specific effector proteins to exploit host cell machinery and hijack the immune responses during infection. Dedicated multiprotein complexes, known as secretion systems, secrete these effectors. Type IV secretion systems (T4SS) are specialized adenosine triphosphate-dependent protein complexes used by many bacterial pathogens for the delivery of Type IV Effectors (T4Es) proteins into eukaryotic cells to subvert host cell processes during infection. To help biologists to identify putative T4Es from bacterial genomes, S4TE2.0 (Meyer *et al.*, 2013; Noroy

Table 1. Selected databases and resources available in the network

Name	Description	Link
SITVITWEB	MTBC genotyping database containing information on 62 582 isolates	http://www.pasteur-guadeloupe.fr:8081/SITVIT_ONLINE/
SITVIT2	An update of SITVITWEB containing additional molecular markers and information on 111 635 isolates	http://www.pasteur-guadeloupe.fr:8081/SITVIT2
SITVITBovis	A publicly available database and mapping tool to get an improved overview of animal and human cases caused by *Mycobacterium bovis* ($n = 25\,741$ isolates)	http://www.pasteur-guadeloupe.fr:8081/SITVIT_Bovis
SpolSimilaritySearch	Similarity search algorithm of spoligotype patterns in the SITVIT2 database	http://www.pasteur-guadeloupe.fr:8081/SpolSimilaritySearch/
SpolLineages	A tool to predict MTBC families from spoligotyping or MIRU-VNTR patterns	http://www.pasteur-guadeloupe.fr:8081/SpolLineages/
S4TE2.0	A tool to predict T4SS effectors from bacterial genomes	https://sate.cirad.fr/S4TE.php
S4TE-EM	A tool based on the S4TE 2.0 algorithm which allows users to adjust all parameters	https://sate.cirad.fr/S4TE-EM.php
S4TE-CG	A tool to compare up to four effectomes predicted by SATE 2.0 algorithm	https://sate.cirad.fr/S4TE-CG.php

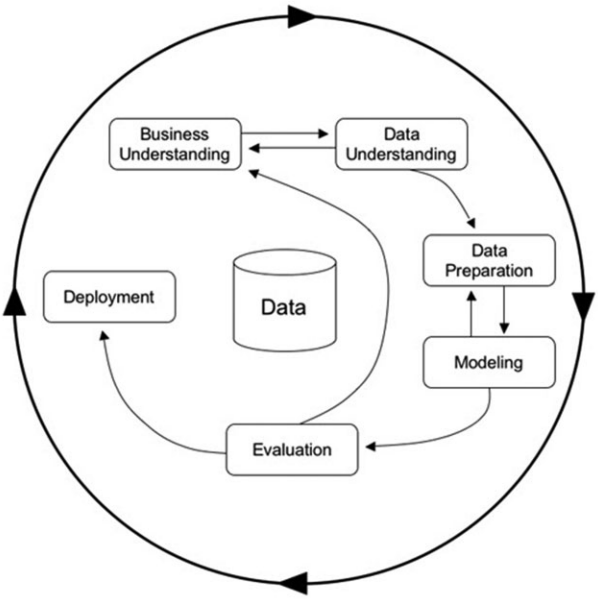

Fig. 3. Cross Industry Standard Process for Data Mining (Crisp-DM) model that describes common approaches used by KDD and data mining experts

et al., 2019; Noroy and Meyer, 2021), an online bioinformatic suite of tools has been developed by researchers from KaruBioNet at CIRAD Guadeloupe. S4TE 2.0 predicts and ranks candidate T4Es by using a combination of 14 sequence characteristics, including homology to known effectors, homology to eukaryotic domains, presence of subcellular localization signals or secretion signals, etc. S4TE 2.0 generates a score and sorts the best T4Es candidates. S4TE-CG allows the comparison of up to four predicted type IV effectomes.

The following table shows various tools and databases which are available for studying specific aspects of infectious diseases (such as *M.tuberculosis* genotyping information or Type IV effector proteins; Table 1).

The links of all these databases developed inside KaruBioNet are available on the website. The development of collaborations will lead to the creation of new databases or to the use of existing databases. The sensitivity of these data (personal data, health data, genetic data, etc.) requires compliance with European data protection rules (GDPR).

7 Machine learning and classification problems

Knowledge Discovery from Databases (KDD) is the 'non-trivial process of identifying valid, novel, potentially useful, and ultimately understandable patterns in data' (Fayyad *et al.*, 1996). Over the last 30 years, advanced prediction methods and tools borrowed to the KDD field, have contributed to defining new kinds of prediction models, namely data-driven models. They consist of looking for correlations between predictive historical variables and output variables. When the output variables are discrete, the problems considered are referred to as 'classification problems'. Figure 3 presents the different steps of a KDD approach. First, it is important to understand the data and the associated context, then a preparation step (also known as preprocessing) is often necessary in order to overcome problems such as missing data, normalization issues, noise reduction or data transformation, and obtain usable data to feed a dedicated model. Among the current models used to tackle these problems, artificial neural networks (ANNs), decision trees or support vector machines are the most common. They have proved to be very effective compared to knowledge-driven solutions with the advantage of requiring less domain knowledge for their design. However, one of the issues in data-driven approaches is the understandability and readability of models that end-users should trust. Indeed, a lot of techniques, including ANNs, can be seen as delivering black-boxes since they do not provide explanations of how they work. Decision-makers are more likely to trust models whose predictions are interpretable and understandable.

8 Conclusion and perspectives

Through this network, we aim to forge links of solidarity allowing a better collaboration between researchers from various fields. Concerted, collaborative and coordinated actions are still needed to better structure bioinformatics in our environment. Indeed, several examples of well-established bigger structures at national level already exist such as the French Institute of Bioinformatics (https://www.france-bioinformatique.fr/en), the Center of Bioinformatics Biostatistics and Integrative Biology (https://c3bi.pasteur.fr/) or the South Green Bioinformatics platform (https://www.southgreen.fr/; South Green Collaborators, 2016). Other examples of structuring are visible at a wider scale. Furthermore, educational initiatives such as Meet-U (http://www.meet-u.org/) are inspiring to set up innovative and original learning processes and collaboration (Abdollahi *et al.*, 2018). Starting from a little archipelago in the Caribbean, we envision to facilitate exchange with other Caribbean partners in the future, knowing that the Caribbean is a region with a non-negligible biodiversity, which should be preserved and protected. Joining forces with other 'isolated' Caribbean countries or territories would strengthen scientific collaborative impact in the region (Degrave *et al.*, 2002; Gaete and Deloumeaux, 2017). Finally, the KaruBioNet network could potentially evolve by collaborating with other Caribbean countries, and give rise to another greater network tentatively called KariBioNet. Further efforts are needed to achieve this goal. However, the actions initiated by the launch of an open-access Galaxy instance and a website bringing together our activities could help many users in the analysis of their data.

Acknowledgements

The KaruBioNet network thanks the Regional Council of Guadeloupe for its support. We are also thankful to the 'Projet MALIN' consortium (https://www.projet-malin.fr/). This work was supported by a FEDER grant financed by the EU. Several computational tests have been performed using Wahoo and Exocet, the clusters of the Centre Commun de Calcul Intensif (C3I) of the Université des Antilles. We are grateful to Nalin Rastogi for helpful discussions. We also thank other people interested or involved in the KaruBioNet team: Raphaël Pasquier, Syndia Sadikalay, Pauline Dentika, Anubis Vega-Rúa, Lyza Hery, Margaux Mulatier, Elodie Calvez, Gélixa Gamiette, Antoine Talarmin, Stéphanie Guyomard, Degrâce Batantou, Vincent Guerlais, Maïlie Saint-Hilaire, Cécile Martias, Thierry Zozio, Isaure Quétel, Gaëlle Gruel, Nina Allouch, Youri Vingataramin, Marc Romana, Michel Naves, Kizzy-Clara Cita, Daniella Goindin, Alice Choury, Olivier Gros, Larissa Valmy, Jean-Christophe Bambou, Antoine Boullis, Manuel Clergue, Sébastien Regis, Denis Boucaud-Maitre, Christophe Armand, Steve Cériac, Yann Legros, Hugo Boijout, Jean-David Pommier, Jimmy Nagau, Andrei Doncescu, Alain Piétrus, Stéphane Cholet, Jean-Luc Gourdine, Tenissia Cesar, Christopher Cambrone, Georges Minatchy, Suzanne Conjard, Carole Louis-Rose, Elvire Couchy, Murielle Mantran, Vincent Moco, Davy Régalade, Elkana Lesmond, Suly Rambinaising, Mame-Boucar Diouf, Sylvaine Bastian, Raymond Césaire, Benoit Garin, Cécile Herrmann, Chantal Eucar, Charlotte Romero, Kévin Julianus, Kévin Durimel, Lucie Leon, Mounir Serag, Muriel Nicolas, Annthommy Gilles, Dominique Albina, Alex Botino, Betty Fausta, Olivier Watté, Jean-François Dorville, Audrey Robinel, Kenny Chammougon and Éric Tessanne. This initiative was presented at the Caribbean Science and Innovation Meeting in 2019. We also thank Fabien Mareuil, Rémi Planel and Brice Raffestin (Institut Pasteur, Paris), for their help regarding the development of Galaxy. Finally, we are grateful to all present and future partners interested in this project.

Funding

This study was partly conducted in the framework of the project MALIN 'Surveillance, diagnosis, control and impact of infectious diseases of humans,

animals and plants in tropical islands', grant # 2015-FED-186, supported by the European Union in the framework of the European Regional Development Fund (ERDF) and the Regional Council of Guadeloupe. The funders had no role in study design, data collection and analysis, decision to publish or preparation of the manuscript.

Conflict of Interest: none declared.

References

Abdollahi,N. *et al.* (2018) Meet-U: educating through research immersion. *PLoS Comput. Biol.*, **14**, e1005992.

Afgan,E. *et al.* (2018) The Galaxy platform for accessible, reproducible and collaborative biomedical analyses: 2018 update. *Nucleic Acids Res.*, **46**, W537–W544.

Alonge,M. *et al.* (2019) RaGOO: fast and accurate reference-guided scaffolding of draft genomes. *Genome Biol.*, **20**, 224.

Bankevich,A. *et al.* (2012) SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. *J. Comput. Biol.*, **19**, 455–477.

Carattoli,A. *et al.* (2014) In silico detection and typing of plasmids using PlasmidFinder and plasmid multilocus sequence typing. *Antimicrob. Agents Chemother.*, **58**, 3895–3903.

Cohen,K.A. *et al.* (2015) Evolution of extensively drug-resistant tuberculosis over four decades: whole genome sequencing and dating analysis of *Mycobacterium tuberculosis* isolates from KwaZulu-Natal. *PLoS Med.*, **12**, e1001880.

Couvin,D. *et al.* (2017) SpolSimilaritySearch—a web tool to compare and search similarities between spoligotypes of *Mycobacterium tuberculosis* complex. *Tuberculosis (Edinb)*, **105**, 49–52.

Couvin,D. *et al.* (2019) Macro-geographical specificities of the prevailing tuberculosis epidemic as seen through SITVIT2, an updated version of the *Mycobacterium tuberculosis* genotyping database. *Infect. Genet. Evol.*, **72**, 31–43.

Couvin,D. *et al.* (2020) Novel methods included in SpolLineages tool for fast and precise prediction of *Mycobacterium tuberculosis* complex spoligotype families. *Database (Oxford)*, **2020**, baaa108.

Couvin,D. *et al.* (2021) MTBCtools: a non-exhaustive list of software tools/resources for bioinformatics analyses of *Mycobacterium tuberculosis* complex, the causative agent of tuberculosis. *Int. J. Mycobacteriol.*, **9**, 18.

Couvin,D. *et al.* (2022) SITVITBovis—a publicly available database and mapping tool to get an improved overview of animal and human cases caused by *Mycobacterium bovis*. *Database (Oxford)*, **2022**, baab081.

D'Elia,D. *et al.* (2009) The 20th anniversary of EMBnet: 20 years of bioinformatics for the Life Sciences community. *BMC Bioinformatics*, **10 (Suppl 6)**, S1.

Degrave,W.M. *et al.* (2002) Towards a bioinformatics network for Latin America and the Caribbean (LACBioNet). *Appl. Bioinformatics*, **1**, 53–56.

Demay,C. *et al.* (2012) SITVITWEB—a publicly available international multimarker database for studying *Mycobacterium tuberculosis* genetic diversity and molecular epidemiology. *Infect. Genet. Evol.*, **12**, 755–766.

Dereeper,A. *et al.* (2008) Phylogeny.fr: robust phylogenetic analysis for the non-specialist. *Nucleic Acids Res.*, **36**, W465–W469.

Escudié,F. *et al.* (2018) FROGS: find, Rapidly, OTUs with Galaxy Solution. *Bioinformatics*, **34**, 1287–1294.

Fayyad,U.M. *et al.* (1996) Knowledge discovery and data mining: towards a unifying framework. KDD-96, 82–88. https://www.aaai.org/Papers/KDD/1996/KDD96-014.pdf. (30 March 2021, date last accessed).

Gaete,S. and Deloumeaux,J. (2017) Building a network of biological resource centers for research purposes in the Caribbean: excellent potential for research into public health diseases. *Environ. Ecol. Res.*, **5**, 495–499.

Galperin,M.Y. *et al.* (2021) COG database update: focus on microbial diversity, model organisms, and widespread pathogens. *Nucleic Acids Res.*, **49**, D274–D281.

Gruel,G. *et al.* (2021) Critical evaluation of cross-sectoral collaborations to inform the implementation of the 'One Health' approach in Guadeloupe. *Front. Public Health*, **9**, 652079.

Guindon,S. *et al.* (2010) New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. *Syst. Biol.*, **59**, 307–321.

Hazbón,M.H. *et al.* (2018) Mycobacterial biomaterials and resources for researchers. *Pathog. Dis.*, **76**, fty042.

Hunt,M. *et al.* (2019) Antibiotic resistance prediction for *Mycobacterium tuberculosis* from genome sequence data with Mykrobe. *Wellcome Open Res.*, **4**, 191.

Jolley,K.A. *et al.* (2018) Open-access bacterial population genomics: BIGSdb software, the PubMLST.org website and their applications. *Wellcome Open Res.*, **3**, 124.

Katoh,K. and Standley,D.M. (2013) MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Mol. Biol. Evol.*, **30**, 772–780.

Koboldt,D.C. *et al.* (2012) VarScan 2: somatic mutation and copy number alteration discovery in cancer by exome sequencing. *Genome Res.*, **22**, 568–576.

Krzywinski,M. *et al.* (2009) Circos: an information aesthetic for comparative genomics. *Genome Res.*, **19**, 1639–1645.

Lemoine,F. *et al.* (2019) NGPhylogeny.fr: new generation phylogenetic services for non-specialists. *Nucleic Acids Res.*, **47**, W260–W265.

Lerner,H. and Berg,C. (2017) A comparison of three holistic approaches to health: One Health, EcoHealth, and Planetary Health. *Front. Vet. Sci.*, **4**, 163.

Letunic,I. and Bork,P. (2021) Interactive Tree Of Life (iTOL) v5: an online tool for phylogenetic tree display and annotation. *Nucleic Acids Res.*, **49**, W293–W296.

Li,H. (2014) Toward better understanding of artifacts in variant calling from high-coverage samples. *Bioinformatics*, **30**, 2843–2851.

Li,H. *et al.*; 1000 Genome Project Data Processing Subgroup. (2009) The Sequence Alignment/Map format and SAMtools. *Bioinformatics*, **25**, 2078–2079.

Liu,B. *et al.* (2019) VFDB 2019: a comparative pathogenomic platform with an interactive web interface. *Nucleic Acids Res.*, **47**, D687–D692.

Meyer,D.F. *et al.* (2013) Searching algorithm for type IV secretion system effectors 1.0: a tool for predicting type IV effectors and exploring their genomic context. *Nucleic Acids Res.*, **41**, 9218–9229.

Mulder,N.J. *et al.*; H3ABioNet Consortium. (2016) H3ABioNet, a sustainable pan-African bioinformatics network for human heredity and health in Africa. *Genome Res.*, **26**, 271–277.

Noroy,C. and Meyer,D.F. (2021) The super repertoire of type IV effectors in the pangenome of Ehrlichia spp. provides insights into host-specificity and pathogenesis. *PLoS Comput. Biol.*, **17**, e1008788.

Noroy,C. *et al.* (2019) Searching algorithm for Type IV effector proteins (S4TE) 2.0: improved tools for Type IV effector prediction, analysis and comparison in proteobacteria. *PLoS Comput. Biol.*, **15**, e1006847.

Page,A.J. *et al.* (2015) Roary: rapid large-scale prokaryote pan genome analysis. *Bioinformatics*, **31**, 3691–3693.

Page,A.J. *et al.* (2020) Rapid *Mycobacterium tuberculosis* spoligotyping from uncorrected long reads using Galru. *bioRxiv* 2020.05.31.126490. doi: 10.1101/2020.05.31.126490

Phelan,J.E. *et al.* (2019) Integrating informatics tools and portable sequencing technology for rapid detection of resistance to anti-tuberculous drugs. *Genome Med.*, **11**, 41.

Pot,M. *et al.* (2021) Wide distribution and specific resistance pattern to third-generation Cephalosporins of *Enterobacter cloacae* complex members in humans and in the environment in Guadeloupe (French West Indies). *Front. Microbiol.*, **12**, 628058.

Price,M.N. *et al.* (2010) FastTree 2—approximately maximum-likelihood trees for large alignments. *PLoS One*, **5**, e9490.

Robertson,J. and Nash,J.H.E. (2018) MOB-suite: software tools for clustering, reconstruction and typing of plasmids from draft assemblies. *Microb. Genom.*, **4**, e000206.

Seemann,T. (2014) Prokka: rapid prokaryotic genome annotation. *Bioinformatics*, **30**, 2068–2069.

South Green Collaborators. (2016) The South Green portal: a comprehensive resource for tropical and Mediterranean crop genomics. *Curr. Plant Biol.*, **7**, 6–9.

Stamatakis,A. (2014) RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics*, **30**, 1312–1313.

Tang,C.Y. and Ong,R.T. (2020) MIRUReader: MIRU-VNTR typing directly from long sequencing reads. *Bioinformatics*, **36**, 1625–1626.

Volant,S. *et al.* (2020) SHAMAN: a user-friendly website for metataxonomic analysis from raw reads to statistical analysis. *BMC Bioinformatics*, **21**, 345.

Wick,R.R. *et al.* (2017) Unicycler: resolving bacterial genome assemblies from short and long sequencing reads. *PLoS Comput. Biol.*, **13**, e1005595.

Xia,E. *et al.* (2016) SpoTyping: fast and accurate in silico Mycobacterium spoligotyping from sequence reads. *Genome Med.*, **8**, 19.

Zankari,E. *et al.* (2012) Identification of acquired antimicrobial resistance genes. *J. Antimicrob. Chemother.*, **67**, 2640–2644.