Food-Microbiomes Transfert, a shotgun metagenomic tool and a database to analyze cheese ecosystems

Thibaut Guirimand, Anne-Laure Abraham, Sandra Derozier, Charlie Pauvert, Mahendra Mariadassou, Valentin Loux, Pierre Renault

Food-Microbiomes Transfert, a shotgun metagenomic tool and a database to analyze cheese ecosystems

Thibaut GUIRIMAND[1], Anne-Laure ABRAHAM[2], Sandra DEROZIER[2], Charlie PAUVERT[3], Mahendra MARIADASSOU[2], Valentin LOUX[2], and Pierre RENAULT[1]

[1] Micalis Institute, INRA, AgroParisTech, Université Paris-Saclay, 78350 Jouy-en-Josas, France

[2] MaIAGE, INRA, Université Paris-Saclay, 78350 Jouy-en-Josas, France

[3] BIOGECO, INRA, Université de Bordeaux, 69 Route d'Arcachon, 33612 Cestas, France

Corresponding Author: thibaut.guirimand@inra.fr

Abstract *The manufacturing process of cheeses, as for most fermented food, involves a complex flora, composed of bacteria but also yeast and filamentous fungi. These organisms can be brought by starters added during cheese manufacturing, or by environment (milk, maturing cellar…). The exact composition of cheeses is not known. Both academic researchers and cheese manufacturers are interested to have a better insight of cheese ecosystems, and collaborate in the Food-Microbiomes Transfert project. One of the objectives was to develop a metagenomic approach with a user-friendly tool, adapted to cheese ecosystems.*

Keywords *Metagenomics, next generation sequencing, microbial genomes, Database, Web server*

1 Cheese ecosystems characterization with a metagenomic approach

The manufacturing process of cheeses, as for most fermented food, involves a complex flora, composed of bacteria but also yeast and filamentous fungi. The wide range of final products found on the dairy market is representative of the diversity of natural starters and ripening cultures used by dairy industries or coming from the food chain, from milk to the factory. However the cheese ecosystem is not completely understood [1]. The natural starters are not constructed from pure strains and the knowledge of their exact composition remains incomplete. Classical microbiological analysis or genetic methods (qPCR, MLST …) can be used to better understand cheese ecosystem, but these techniques are expensive and time consuming. In order to further understand cheeses ecosystems and maintain a constant quality of cheese products, there is a need for a method to characterize low abundant species and assign precisely taxonomy, in cheese samples.

Techniques based on metagenomic DNA sequencing have been developed recently to rapidly identify species in complex ecosystems. Several tools are available to manage shotgun sequencing metagenomic datasets. Some are based on marker genes (for example: MetaPhlaAn [2], MetaPhyler [3], mOTU [4]) and propose rapid approach to identify species, although the use of a small part of the genomes decreases sensitivity of these approaches. Others use different strategies to take into account all the reads, for example with k-mer approaches (CLARK [5], Kraken [6], LMAT [7], OneCodex [8]...), read mapping on reference genomes (Genometa [9], GOTTCHA [10], MEGAN [11], MicrobeGPS [12], Sigma [13]...), assembly

and functional annotation (EBI metagenomic web server [14], MG-RAST [15]...). Very few are available for biologists.

We are working in partnership with dairy manufacturers to develop a metagenomic approach based on shotgun sequencing of the cheese samples adapted to cheese ecosystems. Cheese ecosystem contain a reduced number of species (less than one hundred in most of the cases) and lots of reference genomes have been sequenced. However, there is a need to assign taxonomy of the present organisms, up to the strain level if possible, and to identify low abundance species. As different strains can have different property on the cheese manufacturing, it is important to have a tool able to identify genes present in the ecosystems. We have developed a method based on the mapping of metagenomic reads on a set of reference genomes, completed by the analysis of reads distributions to identify present species. We also provide information on genes coverage and functions. We have implemented this analysis method under a python pipeline named GeDI.

2 Food-Microbiomes Transfert, a specific database and an interface to analyze cheese ecosystems

Food-Microbiomes Transfert aims to provide a user-friendly tool to analyze cheeses ecosystems. To create the most accessible tool, the project offers a web interface to GeDI and a cheese specific genomes database. This interface will allow users to analyze their own metagenomes (or public metagenomes).

The genomes database has been created using PostgreSQL and currently contains cheese specific microorganisms. These genomes have been extracted from public databases by dairy products ecosystems experts and will be enriched with ecological metadata using text-mining tools and the Ontobiotope ontology [16]. In addition to these public genomes, the user is able to add his own private genomes to perform analysis.
A metagenome database allows to store metagenomics raw data and GeDI results. For this purpose, a metadata model representing cheese manufacturing, sampling method and cheese classification has been created, in partnership with cheese manufacturers and researchers to identify the most accurate and accessible model.
These two parts of the database have been conceived using the Minimum Information about a Genome/Metagenome Sequence (MIGS and MIMS [17]). It allows to have a standard and reusable set of data.

The client side is developed using JavaScript/HTML5/CSS3/RDFa and interacts with the server using AJAX queries. The aim is to create a dynamic interface with minimal user interaction needs and an easy way to perform/manage analysis and data.
The user can upload data, share them with other users, manage genomes lists to reuse for a later analysis, and perform analysis with a minimal steps amount: i) select the metagenome ii) select genomes from public or private lists iii) start GeDI with default parameters.
A results page allows the user to visualize the mapping summary and to download the results (mapping tables, charts).

The server, hosted by the Migale platform, is based on two specific technologies. The web server uses the Python Django framework to manage web client requests, databases and users. The GeDI computation is done on a cluster using a Galaxy [18] instance called by the Python Bioblend library [19]. The use of Galaxy facilitate the reproducibility of research because of the

possibility of exporting the histories and tools. It also allows to easily link the web server to the analysis pipeline because of the use of the same language: Python.

3 Prospectives

We are working on the improvement of GeDI tool: validation on several datasets, computation time... The genome database will be enriched with new genomes and expert annotations especially with text-mining tools. The metadata of the metagenomic database will be added. We are also working on the interface improvement in order to make analysis even more intuitive, and to provide tools to perform cross-comparisons between analyses.

Acknowledgements

We are grateful to the INRA Migale bioinformatics platform (http://migale.jouy.inra.fr) for providing help and support. We also would like to thanks Claire Nédellec, Robert Bossy and Estelle Chaix from the INRA Bibliome team for their work and their support to use Ontobiotope.

References

[1] Almeida M, Hébert A, Abraham AL, Rasmussen S, Monnet C, Pons N, Delbès C, Loux V, Batto JM, Leonard P, Kennedy S. *Construction of a dairy microbial genome catalog opens new perspectives for the metagenomic analysis of dairy fermented products*. BMC genomics, 13;15(1):1101, 2014 Dec

[2] Segata N, Waldron L, Ballarini A, Narasimhan V, Jousson O, Huttenhower C. Metagenomic microbial community profiling using unique clade-specific marker genes. Nature methods, 1;9(8):811-4, 2012 Aug

[3] Liu B, Gibbons T, Ghodsi M, Pop M. MetaPhyler: *Taxonomic profiling for metagenomic sequences. InBioinformatics and Biomedicine (BIBM*), pages 95-100, 2010 IEEE International Conference, 2010 Dec 18

[4] Sunagawa S, Mende DR, Zeller G, Izquierdo-Carrasco F, Berger SA, Kultima JR, Coelho LP, Arumugam M, Tap J, Nielsen HB, Rasmussen S. *Metagenomic species profiling using universal phylogenetic marker genes*. Nature methods, Dec 1;10(12):1196-9, 2013.

[5] Ounit R, Wanamaker S, Close TJ, Lonardi S. *CLARK: fast and accurate classification of metagenomic and genomic sequences using discriminative k-mers.* BMC genomics, 25;16(1):236, 2015 Mar

[6] Wood DE, Salzberg SL. *Kraken: ultrafast metagenomic sequence classification using exact alignments.* Genome biology, Mar 3;15(3):R46, 2014

[7] Ames SK, Hysom DA, Gardner SN, Lloyd GS, Gokhale MB, Allen JE. *Scalable metagenomic taxonomy classification using a reference genome database.* Bioinformatics, 15;29(18):2253-60, 2013 Sep

[8] Minot SS, Krumm N, Greenfield NB. *One Codex: A Sensitive and Accurate Data Platform for Genomic Microbial Identification.* bioRxiv, 1:027607, 2015 Jan

[9] Davenport CF, Neugebauer J, Beckmann N, Friedrich B, Kameri B, Kokott S, Paetow M, Siekmann B, Wieding-Drewes M, Wienhöfer M, Wolf S. *Genometa-a fast and accurate classifier for short metagenomic shotgun reads*. PloS one, 21;7(8):e41224, 2012 Aug.

[10] Freitas TA, Li PE, Scholz MB, Chain PS. *Accurate read-based metagenome characterization using a hierarchical suite of unique signatures.* Nucleic acids research, 12:gkv180, 2015 Mar

[11] Huson DH, Auch AF, Qi J, Schuster SC. *MEGAN analysis of metagenomic data.* Genome research, 1;17(3):377-86, 2007 Mar

[12] Lindner MS, Renard BY. *Metagenomic profiling of known and unknown microbes with MicrobeGPS*. PloS one, 2;10(2):e0117711, 2015 Feb

[13] Ahn TH, Chai J, Pan C. *Sigma: Strain-level inference of genomes from metagenomic analysis for biosurveillance.* Bioinformatics, 29:btu641, 2014 Sep

[14] Hunter S, Corbett M, Denise H, Fraser M, Gonzalez-Beltran A, Hunter C, Jones P, Leinonen R, McAnulla C, Maguire E, Maslen J. *EBI metagenomics—a new resource for the analysis and archiving of metagenomic data*. Nucleic acids research, 1;42(D1):D600-6, 2014 Jan

[15] Meyer F, Paarmann D, D'Souza M, Olson R, Glass EM, Kubal M, Paczian T, Rodriguez A, Stevens R, Wilke A, Wilkening J. *The metagenomics RAST server–a public resource for the automatic phylogenetic and functional analysis of metagenomes*. BMC bioinformatics, 19;9(1):386, 2008 Sep

[16] Bossy R, Golik W, Ratkovic Z, Bessières P, Nédellec C. *Bionlp shared task 2013–an overview of the bacteria biotope task*. InProceedings of the BioNLP Shared Task pages 161-169, 2013 Workshop, 2013 Aug

[17] Yilmaz P, Kottmann R, Field D, Knight R, Cole JR, Amaral-Zettler L, Gilbert JA, Karsch-Mizrachi I, Johnston A, Cochrane G, Vaughan R. *Minimum information about a marker gene sequence (MIMARKS) and minimum information about any (x) sequence (MIxS) specifications*. Nature biotechnology, 1;29(5):415-20, 2011 May.

[18] Goecks J, Nekrutenko A, Taylor J. *Galaxy: a comprehensive approach for supporting accessible, reproducible, and transparent computational research in the life sciences*. Genome biology, 25;11(8):R86, 2010 Aug

[19] Sloggett C, Goonasekera N, Afgan E. *BioBlend: automating pipeline analyses within Galaxy and CloudMan*. Bioinformatics, 1;29(13):1685-6, 2013 Jul