Host tropism and host-pathogen interplay of typhoidal Salmonella enterica

Ludovic Mallet, Claire Hoede, Franck Cerruti, Annick Moisan, Christine Gaspin, Isabelle Virlogeux-Payant, Inna Shomer, Ohad Gal-Mor, Thomas Schiex, Helene Chiapello

▶ **To cite this version:**

Ludovic Mallet, Claire Hoede, Franck Cerruti, Annick Moisan, Christine Gaspin, et al.. Host tropism and host-pathogen interplay of typhoidal Salmonella enterica. JOBIM 2017 - Journées Ouvertes en Biologie, Informatique et Mathématiques, Jul 2017, Lille, France. , 2017, Journées Ouvertes en Biologie, Informatique et Mathématiques. JOBIM 2017. hal-01603254

HAL Id: hal-01603254
https://hal.science/hal-01603254

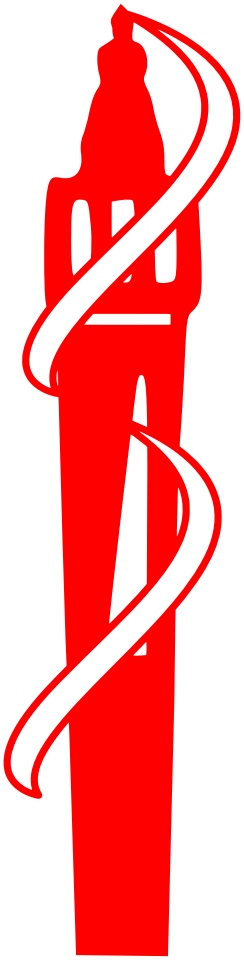
Submitted on 2 Jun 2020

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Distributed under a Creative Commons Attribution - ShareAlike 4.0 International License

Journées Ouvertes en Biologie, Informatique et Mathématiques



JOBIM 2017 Lille

Éditeurs
Cédric Lhoussaine
Hélène Touzet

Lille, 3-6 juillet 2017

Préface

Cette année, pour sa dix-huitième édition, JOBIM (Journées Ouvertes en Biologie, Informatique et Mathématiques) fait étape à Lille. Cet événement annuel, proposé conjointement par la Société Française de BioInformatique, le GDR BioInformatique Moléculaire et l'Institut Français de Bioinformatique, est devenu au fil du temps un rendez-vous incontournable de la communauté bioinformatique et biostatistique nationale. JOBIM offre une photographie unique des travaux menés dans l'Hexagone, et parfois même un peu plus loin. C'est un lieu d'échanges et de partages, toujours convivial, où se croisent chercheurs et ingénieurs, théoriciens et praticiens, jeunes et moins jeunes.

Comme à son habitude, le programme est très riche. Il reflète la diversité et la vitalité de la discipline : analyse de données omiques, phylogénie, biologie des systèmes, bioinformatique structurale, biologie intégrative,... Le comité de programme a sélectionné 37 présentations orales, dont 22 exposés classiques et 15 démonstrations logicielles, et 100 posters. Nous avons également le plaisir d'accueillir six orateurs invités : Céline Brochier-Armanet (LBBE, Lyon), Franca Fraternali (King's college, London), John Huelsenbeck (UC Berkeley), Tobias Marschall (MPI, Saarbrücke), Julio Saez-Rodrigues (EMBL-EBI, Aachen) et Patrick Wincker (Génoscope). Vous trouverez tous les résumés dans ce volume. Enfin, se tiendront en marge de JOBIM les journées de l'association JeBiF (Jeunes Bioinformaticiens de France).

Nous vous souhaitons une belle conférence !

Pour le comité de programme,
Cédric Lhoussaine et Hélène Touzet

Pour le comité d'organisation,
Guillemette Marot et Jean-Stéphane Varré

Comité de programme

Cédric Lhoussaine (CRIStAL, Université de Lille)
Hélène Touzet (CNRS, CRIStAL, Université de Lille)

Eric Bapteste (CNRS, AIRE, Université Pierre et Marie Curie)
Sèverine Bérard (ISEM, Université Montpellier)
Virginie Bernard (Institut Curie, Paris)
Samuel Blanquart (Inria Lille)
Jérémie Bourdon (LINA, Université de Nantes)
Bastien Boussau (CNRS, LBBE, Université Lyon 1)
Christine Brun (CNRS, TAGC, Aix-Marseille Université)
Ségolène Caboche (TAG/CIIL, Université de Lille)
Frédéric Cazals (Inria Sophia)
Annie Chateau (LIRMM, Université Montpellier)
Rayan Chikhi (CNRS, CRIStAL, Université de Lille)
Vincent Chouraki (Inserm U1167)
Sarah Cohen-Boulakia (LRI, Universite Paris-Sud)
Marie-Agnès Dillies (Institut Pasteur, Paris)
Damien Eveillard (LINA, Université de Nantes)
Jérôme Feret (Inria / Ecole Normale Supérieure)
Martin Figeac (Université de Lille)
Christine Gaspin (Inra Toulouse)
Carito Guziolowski (LS2N, Ecole Centrale de Nantes)
Laurent Jacob (CNRS, LBBE, Université Lyon 1)
Fabien Jourdan (INRA Toulouse)
Michaël Kopp (I2M, Aix-Marseille Université)
Vincent Lacroix (LBBE, Université Lyon 1)
Valérie Leclère (Institut Charles Viollette, Université de Lille)
Claire Lemaitre (Inria Rennes)
Marc Lensink (CNRS, UGSF, Université de Lille)
Anne Lopes (IGM, Université Paris Sud)
Macha Nikoslki (CNRS, LaBRI and CBIB, Université de Bordeaux)
Loïc Paulevé (CNRS, LRI, Université Paris-Sud)
Eric Pelletier (CEA / Genoscope)
Guy Perrière (CNRS, LBBE, Université Lyon 1)
Pierre Peterlongo (Inria Rennes)
Pierre Peyret (MEDIS, INRA-Université Clermont Auvergne)
Céline Poux (EEP, Université de Lille)
Hugues Richard (LCQB, Université Pierre et Marie Curie)
Delphine Ropers (Inria Grenoble Rhône-Alpes)
Nicolas Servant (Institut Curie, Paris)
Anne Siegel (CNRS, IRISA, Unversité Rennes 1)
Hédi Soula (Université Pierre et Marie Curie)
Pascal Touzet (EEP, Université de Lille)
Jacques van Helden (TAGC, Aix-Marseille Université)
Vincent Vandewalle (EA 2694, Université de Lille)
Cristian Versari (CRIStAL, Université de Lille)

Comité d'organisation

Marie-Bénédicte Dernoncourt (Inria Lille)
Guillemette Marot (EA 2694, Université de Lille)
Jean-Stéphane Varré (CRIStAL, Université de Lille)

Émilie Allart (CRIStAL, Université de Lille)
Anaïs Barray (bilille, Université de Lille)
Aurélien Béliard (CHRU Lille, CRIStAL, Université de Lille)
Samuel Blanck (EA 2694, Université de Lille)
Ségolène Caboche (TAG/CIIL, Université de Lille)
Hélène Chiapello (Inra Toulouse, SFBI)
Christophe Demay (CHRU Lille)
Gaël Even (Genes Diffusion)
Sophie Gallina (CNRS, EEP, Université de Lille)
Sébastien Grec (UGSF, Université de Lille)
Isabelle Guigon (bilille, Université de Lille)
David Hot (Institut Pasteur de Lille)
Stéphane Janot (CRIStAL, Université de Lille)
Laetitia Jourdan (CRIStAL, Université de Lille)
Valérie Leclère (Institut Charles Viollette, Université de Lille)
Pierre Marijon (Inria Lille)
Laurent Noé (CRIStAL, Université de Lille)
Pierre Pericard (CRIStAL, Université de Lille)
Céline Poux (EEP, Université de Lille)
Maude Pupin (CRIStAL, Université de Lille)
Tatiana Rocher (CRIStAL, Université de Lille)
Chadi Saad (CRIStAL, JPARC, Université de Lille)
Mikaël Salson (CRIStAL, Université de Lille)
Olivier Sand (UMR8199, Institut de Biologie de Lille)
Pascal Touzet (EEP, Université de Lille)

Table des matières

Mardi 4 juillet

Mercredi 5 juillet

Jeudi 6 juillet

Posters 105

Présentations orales

Conférence invitée

Patrick Wincker

CEA Genoscope, Evry, France

Holistic metagenomics in marine plankton communities

Simka: large scale *de novo* comparative metagenomics

Gaëtan BENOIT[1], Pierre PETERLONGO[1], Mahendra MARIADASSOU[2], Erwan DREZEN[1,3], Sophie SCHBATH[2], Dominique LAVENIER[1] and Claire LEMAITRE[1]

[1] INRIA/IRISA, Genscale team, UMR6074 IRISA CNRS/INRIA/Université de Rennes 1, Campus de Beaulieu, 35042, Rennes, France
[2] MaIAGE, INRA, Université Paris-Saclay, 78350, Jouy-en-Josas, France
[3] CHU Pontchaillou, 35000, Rennes, France

Corresponding author: gaetan.benoit@inria.fr

Reference paper: **Benoit *et al.* (2016) Multiple comparative metagenomics using multiset k-mer counting. *PeerJ Computer Science*. https://doi.org/10.7717/peerj-cs.94**

Large metagenomic projects, such as Human Microbiome Project [1] or Tara Ocean Project [2], are becoming increasingly important for understanding life processes at the individual scale or at more global scales. Therefore, serious efforts are put in funding projects, collecting DNA, sequencing, and analyzing the resulting sequences. Comparative metagenomics is one of the most ubiquitous and informative of those analyzes. The purpose is mainly to estimate proximity (or distance) between two or more environmental sites at the genomic level. The comparisons are often based on identified species content. However, this approach is limited to sequences correctly assigned to known species documented in public biobanks, and this may correspond to small fractions of the datasets, in particular for environmental samples such as sea water. This limitation motivated the development of *de novo* comparison tools, such as Compareads [3] or Mash [4], based only on the non assembled read set comparisons. Compareads was for instance successfully used for the first analyses of the Tara Ocean data [5].

These reference-free methods share the use of k-mers as the fundamental unit used for comparing samples. Actually, k-mers are a natural unit for comparing communities: (1) sufficiently long k-mers are usually specific of a genome [6], (2) k-mer frequency is linearly related to genome's abundance [7], (3) k-mer aggregates organisms with very similar k-mer composition (*e.g.* related strains from the same bacterial species) without need for a classification of those organisms [8]. However, even if Compareads approach was designed to scale-up to large metagenomic read sets, its use on data generated by large scale projects is turning into a bottleneck in terms of time requirements. By contrast, Mash outperforms by far all other methods in terms of computational resource usage. However, this frugality comes at the expense of result quality and precision: the output distances and Jaccard indexes do not take into account relative abundance information and are not computed exactly due to k-mer sub-sampling. This is what motivated this work in which we propose a new *de novo* comparative metagenomic method, called Simka. Simka compares N metagenomic datasets based on their exact k-mers counts. It computes a large collection of distances classically used in ecology to compare communities, by replacing species counts by k-mer counts, for a large range of k-mer sizes, including large ones (up to 30). Simka is, to our knowledge, the first method able to rapidly compute a full range of distances enabling the comparison of any number of datasets. Simka outperforms state-of-the-art read comparison methods in terms of computational needs and result quality. For instance, Simka ran on 690 samples from the Human Microbiome Project (HMP) (totalling 32 billion reads) in less than 10 hours and using no more than 70 GB RAM.

Simka works as follows. Firstly, the k-mer spectrum of each dataset is computed. The k-mer spectrum of a dataset is the set of all its distinct k-mers associated with their abundance in the dataset. Secondly, k-mer spectra are compared in a pairwise manner to compute their distance. This comparison process basically aims at identifying which k-mers are shared by both spectra and which ones are not. It can be computationally very expensive because each k-mer spectrum can contains millions to billions of distinct k-mers when k is large (> 15). Moreover, the number of comparisons grows quadratically with the number of input datasets. To tackle this issue, we have designed a new k-mer counting strategy of numerous datasets, called Multiset k-mer Counting (MKC). MKC takes N datasets as input and provides an abundance vector for each distinct k-mer. The abundance vector of a k-mer consists of its N counts in the N datasets. The abundance vector generation by the MKC task is divided into two phases: (1) Sorting Count, (2) Merging Count. During the first step, the k-mers of each dataset are counted independently. This is performed by sorting the k-mers in lexicographical order. Distinct k-mers can thus be identified and their number of occurrences computed. This task can be very efficiently performed by popular disk-based k-mer counting tool such as DSK [9] or KMC2 [10]. The resulting k-mer spectra are written on the disk. During the second step, a Merge-Sort algorithm can be efficiently

applied on the sorted k-mer spectra to directly generate abundance vectors. Given those abundance vectors, the distances between each pair of datasets can be computed simultaneously. Interestingly, most of the ecological distances are additive over the distinct k-mers, meaning that they can be iteratively updated one abundance vector at a time. Once an abundance vector has been processed, there is thus no need to keep it on record, allowing Simka to have a very low memory footprint.

One advantage of the overall Simka workflow is its high parallelism potential. During the sorting count phase of the MKC, a first parallelism level is given by the independent counts of each dataset. N processes can thus be run in parallel, each one dealing with a specific dataset. A second level is given by the fine grained parallelism implemented in software such as DSK or KMC2 that intensively exploit today multicore processor capabilities. As the number of distinct k-mers is generally huge, those tools separate the k-mers in P smaller disjoint sets that can be counted independently and thus result in P k-mer spectrum chunks per dataset. During its second step, the MKC exploit this partitionning to merge up to P k-mer spectrums chucks in parallel. Each of these merge processes generates abundance vectors from which independent contribution to the distances are computed. Since the distance computed by Simka are additive over the distinct k-mers, each contribution is simply accumulated and the final distance is computed. Simka implementation is based on the GATB library [11], a C++ library optimized to handle very large sets of k-mers. Simka is usable on standard computers and has also been entirely parallelized for grid infrastructures made of hundred of nodes, and where each node implements 8 or 16-core systems.

The quality of the distances computed by Simka were evaluated answering two questions. First, are they similar to distances between read sets computed using other *de novo* approaches? Second, do they recover the known biological structure of HMP samples? For the first evaluation, we show that Simka result are perfectly well correlated with Compareads results. We go further in this evaluation by showing that Simka results are highly correlated with costly but extremely accurate *de novo* comparison techniques relying on all-versus-all sequence alignment strategy. For the second evaluation, Simka distances were compared to taxonomic distances that are a traditional way of comparing metagenomic samples. Taxonomic distances are based on sequence assignation to taxons by mapping to reference databases. To compare Simka to such traditional reference-based methods, we used the HMP dataset. One advantage of this dataset is that it has been extensively studied, in particular the microbial communities are relatively well represented in reference databases [1,12]. We show that substituting k-mer counts by species counts gives admittedly different distances but that those distances are biologically relevant as they capture the same underlying biological structure and lead to the same conclusions as those based on taxonomic composition. In particular, Simka was able to retrieve two major biological results. The first one is the segregation of the HMP datasets by body sites. The second one reveals that the organisation of the gut samples is mainly driven by the relative abundances of three bacterial genera, known as enterotypes, and characterized by the relative abundances of a few genera: *Bacteroides*, *Prevotella* and genera from the *Ruminococcaceae* family. In contrast of Simka, Mash performed badly when considering HMP datasets per body site since this tool can only take into account presence/absence information and not relative abundances. As a matter of fact, differences in relative abundances are subtler signals that are often at the heart of interesting biological insights in comparative genomics studies [13,14,15,16,17].

We introduced Simka, a new method for computing a collection of ecological distances, between many large metagenomic datasets, based on their k-mer composition. This was made possible thanks to the Multiset k-mer Counting algorithm (MKC), a new strategy that counts k-mers of numerous datasets with state-of-the-art time, memory and disk performances.

Acknowledgements

This work was supported by the French ANR-14-CE23-0001 Hydrogen Project.

References

[1] Human Microbiome Project Consortium. Structure, function and diversity of the healthy human microbiome. *Nature*, 486(7402):207–214, 2012.

[2] Eric Karsenti, Silvia G. Acinas, Peer Bork, Chris Bowler, Colomban de Vargas, Jeroen Raes, Matthew Sullivan, Detlev Arendt, Francesca Benzoni, Jean Michel Claverie, Mick Follows, Gaby Gorsky, Pascal Hingamp, Daniele Iudicone, Olivier Jaillon, Stefanie Kandels-Lewis, Uros Krzic, Fabrice Not, Hiroyuki Ogata, Stéphane Pesant, Em-

manuel Georges Reynaud, Christian Sardet, Michael E. Sieracki, Sabrina Speich, Didier Velayoudon, Jean Weissenbach, and Patrick Wincker. A holistic approach to marine Eco-systems biology. *PLoS Biology*, 9, 2011.

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

[4] 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 Biol*, 17(1):132, 2016.

[5] E. Villar, G. K. Farrant, M. Follows, L. Garczarek, S. Speich, S. Audic, L. Bittner, B. Blanke, J. R. Brum, C. Brunet, R. Casotti, A. Chase, J. R. Dolan, F. Ortenzio, J.-P. Gattuso, N. Grima, L. Guidi, C. N. Hill, O. Jahn, J.-L. Jamet, H. Le Goff, C. Lepoivre, S. Malviya, E. Pelletier, J.-B. Romagnan, S. Roux, S. Santini, E. Scalco, S. M. Schwenck, A. Tanaka, P. Testor, T. Vannier, F. Vincent, A. Zingone, C. Dimier, M. Picheral, S. Searson, S. Kandels-Lewis, S. G. Acinas, P. Bork, E. Boss, C. de Vargas, G. Gorsky, H. Ogata, S. Pesant, M. B. Sullivan, S. Sunagawa, P. Wincker, E. Karsenti, C. Bowler, F. Not, P. Hingamp, and D. Iudicone. Environmental characteristics of agulhas rings affect interocean plankton transport. *Science*, 348(6237):1261447–1261447, may 2015.

[6] Y. Fofanov, Y. Luo, C. Katili, J. Wang, Y. Belosludtsev, T. Powdrill, C. Belapurkar, V. Fofanov, T.-B. Li, S. Chumakov, and B. M. Pettitt. How independent are the appearances of n-mers in different genomes? *Bioinformatics*, 20(15):2421–2428, apr 2004.

[7] Yu-Wei Wu and Yuzhen Ye. A novel abundance-based algorithm for binning metagenomic sequences using l-tuples. *Journal of Computational Biology*, 18(3):523–534, 2011.

[8] Hanno Teeling, Jost Waldmann, Thierry Lombardot, Margarete Bauer, and Frank O Glöckner. Tetra: a web-service and a stand-alone program for the analysis and comparison of tetranucleotide usage patterns in dna sequences. *BMC bioinformatics*, 5(1):163, 2004.

[9] Guillaume Rizk, Dominique Lavenier, and Rayan Chikhi. DSK: k-mer counting with very low memory usage. *Bioinformatics*, page btt020, 2013.

[10] Sebastian Deorowicz, Marek Kokot, Szymon Grabowski, and Agnieszka Debudaj-Grabysz. KMC 2: Fast and resource-frugal k-mer counting. *Bioinformatics*, 31(10):1569–1576, 2015.

[11] Erwan Drezen, Guillaume Rizk, Rayan Chikhi, Charles Deltel, Claire Lemaitre, Pierre Peterlongo, and Dominique Lavenier. GATB: Genome assembly & analysis tool box. *Bioinformatics*, 30(20):2959–2961, 2014.

[12] Human Microbiome Project Consortium. A framework for human microbiome research. *Nature*, 486(7402):215–221, Jun 2012.

[13] Sébastien Boutin, Simon Y. Graeber, Michael Weitnauer, Jessica Panitz, Mirjam Stahl, Diana Clausznitzer, Lars Kaderali, Gisli Einarsson, Michael M. Tunney, J. Stuart Elborn, Marcus A. Mall, and Alexander H. Dalpke. Comparison of microbiomes from different niches of upper and lower airways in children and adolescents with cystic fibrosis. *PLoS ONE*, 10(1):1–19, 01 2015.

[14] A. Shade, S. E. Jones, J. G. Caporaso, J. Handelsman, R. Knight, N. Fierer, and J. A. Gilbert. Conditionally rare taxa disproportionately contribute to temporal changes in microbial diversity. *mBio*, 5(4):e01371–14–e01371–14, jul 2014.

[15] S. Genitsaris, S. Monchy, E. Viscogliosi, T. Sime-Ngando, S. Ferreira, and U. Christaki. Seasonal variations of marine protist community structure based on taxon-specific traits using the eastern english channel as a model coastal system. *FEMS Microbiology Ecology*, 91(5):fiv034–fiv034, mar 2015.

[16] Suzanne Coveley, Mostafa S Elshahed, and Noha H Youssef. Response of the rare biosphere to environmental stressors in a highly diverse ecosystem (zodletone spring, ok, usa). *PeerJ*, 3:e1182, 2015.

[17] V. Gomez-Alvarez, S. Pfaller, J. G. Pressman, D. G. Wahman, and R. P. Revetta. Resilience of microbial communities in a simulated drinking water distribution system subjected to disturbances: role of conditionally rare taxa and potential implications for antibiotic-resistant bacteria. *Environ. Sci.: Water Res. Technol.*, 2(4):645–657, 2016.

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 (first 35 nucleotides) on a set of reference genomes with Bowtie [16], with 3 mismatches allowed. These parameters have been chosen to allow

detecting microorganism displaying up to 10% divergence with reference genomes. A first filter based on read distribution on CDS allows discarding false positives. We then propose a method to identify among reference genomes of the same species those that are the closest to ecosystem micro-organisms. It is based on the number of CDS detected on each genome, a comparison between observed and expected CDS coverage [17], and informations about mismatches in the alignments. 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 99 cheese specific microorganisms (we expect soon 300 genomes, with at least one genome for all the genus that have been described in dairy products). 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 [18]. 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 [19]). 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 as shown in Fig 1).

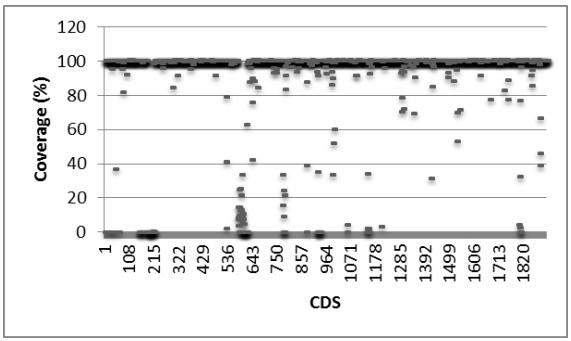


Fig 1. Here is an example of graphics generated by the tool by mapping a 10 million reads metagenome on *Streptococcus thermophilus JIM 8232*. We can observe some uncovered CDS marking differences between the reference genome and the strain present in the ecosystem.

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 [20] instance called by the Python Bioblend library [21]. 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. The actual computation time is about 10 hours for a 1 million reads metagenome and a hundred of reference genomes and about 5 days for a 10 million reads metagenome.

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. GeDI tool will be soon available with command line and Galaxy wrappers. The database and interface will be available via the Migale platform.

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] 16. 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] Langmead B, Trapnell C, Pop M, Salzberg SL.Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. Genome Biology, 2009 Mar.
[17] Lander ES, Waterman MS. Genomic mapping by fingerprinting random clones: A mathematical analysis. Genomics, 2.3, 231-239, 1988.
[18] 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.
[19] 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.
[20] 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.
[21] Sloggett C, Goonasekera N, Afgan E. BioBlend: automating pipeline analyses within Galaxy and CloudMan. Bioinformatics, 1;29(13):1685-6, 2013 Jul.

Reconstruction of full-length 16S rRNA sequences for taxonomic assignment in metagenomics

Pierre PERICARD[1,2], Yoann DUFRESNE[1,2], Samuel BLANQUART[2,1] and Hélène TOUZET[1,2]
[1] CRIStAL (UMR CNRS 9189, Université Lille 1), 59655 Villeneuve d'Ascq Cedex, France
[2] Inria Lille Nord Europe, 59650, Villeneuve d'Ascq, France

Corresponding author: pierre.pericard@gmail.com, helene.touzet@univ-lille1.fr

Abstract *Advances in the sequencing of uncultured environmental samples, raise a growing need for accurate taxonomic assignment. Accurate identification of organisms present within a community is essential to understanding even the most elementary ecosystems. However, current high-throughput sequencing technologies generate short reads which partially cover full-length marker genes and this poses difficult bioinformatic challenges for taxonomy identification at high resolution. We designed MATAM, a software dedicated to the fast and accurate targeted assembly of short reads sequenced from a genomic marker of interest. The method implements a stepwise process based on construction and analysis of a read overlap graph. It is applied to the assembly of 16S rRNA markers and is validated on simulated, synthetic and genuine metagenomes. We show that MATAM outperforms other available methods in terms of low error rates and recovered genome fractions and is suitable to provide improved assemblies for precise taxonomic assignments.*

Keywords Metagenomics, 16S rRNA, assembly, taxonomic assignment, algorithm.

1 Introduction

Shotgun metagenomic sequencing provides an unprecedented opportunity to study uncultured microbial samples, with multiple applications ranging from the human microbiome to soil or marine samples, for which the vast majority of microorganisms diversity remains unknown [1].

A major goal of metagenomic studies is to characterize the microbial diversity and ecological structure. This is often achieved by focusing on one of several phylogenetic marker genes [2,3], that are ubiquitous in the taxonomic range of interest and exhibit variable discriminative regions. For bacterial communities, the gold standard marker is the 16S ribosomal RNA (rRNA, ∼1500bp avg. length), for which millions of sequences are available in curated reference databases, such as Silva [4], RDP [5] or GreenGenes [6]. Traditionnal approaches such as amplicon sequencing are limited to the analysis of small portions of the marker sequences. This leads to strong technological limitations for organisms identification at sufficiently precise taxonomic levels, typically beyond genus [7]. To assign marker sequences to species, or even strains, we need to be able to recover full length rRNA with less than a few errors per kilobase. Metagenomic assemblers are not suitable for this task, because they are optimized to deal with whole genomes, and struggle to differentiate between very similar sequences [8]. To this respect, marker-oriented methods such as EMIRGE [9] and REAGO [10] were recently developed in order to assemble metagenomic read subsets into full length 16S rRNA contigs, thus aiming to improve the taxonomic assignment accuracy of environmental samples. EMIRGE uses a Bayesian approach to iteratively reconstruct 16S rRNA full length sequences. REAGO identifies rRNA reads using Infernal [11], and then constructs an overlap graph by searching for exact overlaps between reads using a suffix/prefix array. However, such tools still show some limitations in terms of recovery error rates as well as dealing with low abundance species.

In this work, we present MATAM, a new approach based on the construction and exploitation of an overlap graph, carefully designed to minimize the error rate and the risk of chimera formation. MATAM was validated on both simulated and actual sequencing data. It is able to reconstruct nearly full length 16S rRNAs and is robust to variations in the sequencing depth as well as community complexity.

2 Methods

2.1 Overview of MATAM

The MATAM (Mapping-Assisted Targeted-Assembly for Metagenomics) pipeline takes as input a set of shotgun metagenomics short reads and a reference database containing the largest possible set of sequences

from a given target marker gene. MATAM identifies reads originating from that marker, and assembles nearly full length sequences of it. It is composed of four major steps. Although this method should work for any conserved and widely surveyed gene, we will focus on the 16S rRNA for the remainder of the article.

2.2 Reference database construction

The availability of a reference database for the marker gene is an essential feature of the method, because it allows us to model the target sequences. For applications to 16S rRNA assembly, MATAM utilizes Silva 128 SSU Ref NR database [4]. From this reference database that we denote as *complete*, we also build a *clustered* reference database, that provides a coarse-grained representation of the taxonomic space. For that task, we use Sumaclust [12,13] (http://metabarcoding.org/sumaclust) using a 95% identity threshold.

2.3 rRNA reads identification and mapping

In the first step, reads are mapped against the clustered reference database using SortMeRNA [14,15]. This step allows to quickly sort out 16S rRNA reads from the whole set of reads, providing high quality alignments. For each read, we keep up to ten best alignments against the reference database. Moreover, this mapping step yields a broad classification of the 16S rRNA reads. Indeed, reads coming from distantly related species are aligned against their respective closest known references, which nest in distant lineages of the taxonomy, while reads from closely related species are aligned against closely related references.

2.4 Construction of the overlap graph

The identified 16S rRNA reads are then organized into an *overlap graph* defined as follows: graph nodes are reads, and an undirected edge connects two nodes if the two reads overlap with a sufficient length and with a sufficient identity to assert that they originated from a common sampled taxon. The standard approach to build such an overlap graph requires comparison of each read with each other, which is time-consuming. Here, we use alignment information to sort through candidate read pairs in a very efficient manner. For each pairing, we consider only reads that share alignments with at least one common reference sequence and for which the alignments are overlapping on more than 50 nucleotides with 100% identity. This strict criterion allows us to reduce the risk of connecting reads from unrelated taxa, which would in turn produce chimeras. By doing so, we discard reads containing sequencing errors in their overlap, which is bearable considering the nowadays very low sequencing error rates of short reads.

2.5 Extracting contigs from the overlap graph

Although the overlap graph appears very bushy, it also reveals some general trends. While it exhibits highly connected subgraphs, it also displays disjoint paths. We simplify the graph by performing a breadth first traversal starting from a random node to annotate the nodes with their depth. All nodes with equal depth that are connected in a single connected component are collapsed into a single *compressed node* and outgoing edges are merged into a *compressed edge*. Low support compressed nodes containing a single read, and compressed edges representing a single overlap are removed. The resulting graph, called the *compressed graph*, is several order of magnitude smaller than the initial overlap graph. We partition this graph in three categories of subgraphs: *hubs*, that are nodes with an degree strictly greater than two, *specific paths* that are sequences of nodes of degree two or one, and *singletons* that are non-connected nodes. Intuitively, hubs correspond to the highly connected subgraphs in the overlap graph, and are likely to contain mainly reads coming from conserved regions shared in many species, thus overlaping without error even for distantly related taxa. Specific paths tend to contain reads originating from variable regions of the 16S gene, that are specific to one or few closely related species. For each subgraph in the compressed graph (hubs, specific paths, singletons), we extract the underlying sets of reads and build an individual assembly using the genomic assembler SGA [16]. Note that any other state-of-the art genomic assembler could be used here. As a result, we obtain one or more contigs for each subgraph.

2.6 Contigs scaffolding

We use a greedy algorithm to scaffold the contigs obtained in the previous step. For that task, contigs are first mapped against the complete reference database, and all alignments within the 1% range of suboptimal scores are kept. We then select contigs by increasing number of matches and decreasing lengths. By doing so, a long contig with a unique alignment will be selected for scaffolding before a short contig exhibiting a large number of alignments. Such long contig can be assigned non-ambiguously to a single species, while the short contig with multiple matches rather corresponds to a conserved region of the marker and is used to fill in the blanks between the specific contigs. Contigs matching against the same reference sequence are then

	Chimera (%)		TAL/TL (%)		ER (%)		Ns (%)		ACL	
	mean	SD	mean	SD	mean	SD	mean	SD	mean	SD
MATAM	1.28	0.55	99.3	0.2	0.03	0.02	0.00	0.00	1252	116.9
EMIRGE	36.89	9.42	79.9	11.6	0.62	0.16	0.55	0.36	1436	15.4
REAGO	42.11	10.36	91.5	0.8	0.31	0.13	0.00	0.00	1333	298.9
SPAdes	21.23	9.05	73.5	15.9	0.60	0.49	0.02	0.04	966	47.4
MEGAHIT	23.81	2.85	80.3	4.9	0.36	0.18	0.00	0.00	962	87.6

Tab. 1. Results for the simulated dataset with varying sequencing depth. We provide averaged metrics for the five sequencing depths. ACL is the average contig length.

merged into a single consensus scaffold. Redundant scaffolds included in larger ones are removed. Finally, only scaffolds larger than 500bp are retained. This yields the final MATAM output which could be used for the purpose of taxonomic assignment.

3 Implementation

MATAM was implemented in Python 3, except for the overlap graph building and compression steps that were written in C++11 using the SeqAn library [17], and is available *via* Docker and Conda. MATAM is distributed under the GNU Affero GPL v3.0 licence and the source code is freely available at the following URL: https://github.com/bonsai-team/matam. All MATAM runs presented in this article were performed using MATAM v0.9.9.

4 Results

MATAM performance was compared with those of two general-purpose metagenomic assemblers, SPAdes [18] and MEGAHIT [19], as well as with two methods specialized in 16S rRNA assembly, EMIRGE [9] and REAGO [10]. The five tools were run on three different datasets, chosen for their complementarity and the possibility to validate the reconstructed candidate 16S rRNA sequences: a simulated dataset [20], a synthetic microbial community [21], and two environmental samples from human gut and mouth providing amplicon based taxonomic assignments [22]. SortMeRNA was used to extract 16S rRNA reads from these datasets before assembling them with SPAdes and MEGAHIT. Complete command-lines and parameters are available in the Supplementary Results.

In order to compare the five methods on a common ground, the same validation procedure was applied for all experiments. Only reconstructed sequences with lengths exceeding 500bp were considered, and chimeric sequences were filtered out by the UCHIME algorithm [23] implemented in VSEARCH [24] and querying the Silva 128 SSU Ref Nr99 database. For each experiment, we indicate the proportion of chimeric contigs (% *chimeras*, which is the total size of all chimeric contigs divided by the assembly total size). All the following measures were then computed on the remaining assemblies. When the sequences present in the sample are actually known (see Sections 4.1 and 4.2), the assembly quality assessment was performed with MetaQuast [25] by aligning the contigs against the original sample sequences, and considering the following metrics: the *number of contigs* (#contigs), which is the total number of contigs of lengths greater than 500bp; the *total length* (TL), which is the total number of bases in the contigs; the *total aligned length* (TAL), which is the total number of aligned nucleotides in the contigs; the *genome fraction* (GF), which stands for the total number of nucleotides from the original sample sequences covered with contigs divided by the total size of the sample sequences; the *error rate* (ER), which consists in the percentage of observed mismatches and indels with respect to the closest matched sequence in the original sample. Finaly, taxonomic assignments were carried out with the RDP Classifier [26].

4.1 Simulated metagenomic datasets with varying sequencing depth

In the first experiment, we evaluated the ability of methods to correctly reconstruct the 16S rRNA sequences in the context of low sequencing depth. For that, we used a selection of 122 genomes providing a realistic taxomical diversity [20,27], that contains 287 distinct 16S rRNA copies. We generated five datasets with varying sequencing depths: 50x, 20x, 10x, 5x and 2x per genome. Illumina reads were simulated with the ART simulator [28], using the HiSeq2500 built-in error profile, 101bp read length, and 250bp fragment length with a 30bp standard-deviation (SD). In this simulation, all species are equally distributed, which corresponds to the *high complexity community* introduced in [20].

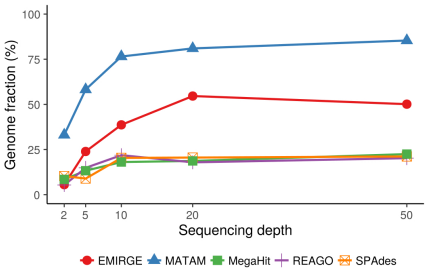


Fig. 1. Effect of sequencing depth on the assemblies genome fractions.

	Chimera (%)	#contigs	TL	TAL	GF (%)	ER (%)	Ns (%)
MATAM	3.2	101	139220	130654	83.1	0.05	0
EMIRGE	17.4	82	117138	102856	50.7	0.17	1.12
REAGO	15.5	59	90269	81297	42.8	0.06	0
SPAdes	5.5	59	70229	59988	39.9	0.11	0.05
MegaHit	3.0	61	77251	68904	44.3	0.18	0

Tab. 2. Results for the synthetic community.

Table 1 shows the results averaged over the five datasets (*mean* metrics and their respective standard deviation, SD). More than 99% of the MATAM sequences were aligned by MetaQuast to one of the 287 16S rRNA sequences from the initial sample (mean TAL/TL), while among other methods, this proportion reached at best 91%, with REAGO. Congruently, MATAM sequences obtained the lowest average error rate (ER=0.03%), which represents more than a ten-fold accuracy gain compared to the other assemblers, and a twenty-fold improvement over EMIRGE. Furthermore, EMIRGE sequences contained 0.5% of unknown nucleotides (Ns), bringing its effective ER above 1%. Additionally, MATAM recovered about thirty times less chimeras than REAGO and EMIRGE did.

For each of the five tools, we reported the recovered genome fraction (GF) with respect to increasing sequencing depth (Figure 1). MATAM recovered from 76% to 85% of the reference sequences for sequencing depths greater than 10x, while EMIRGE recovered less than 55% of the reference sequence, and the GF for other methods is lower than 22%. MATAM also achieved the best performance facing a low sequencing depth of 2x, reaching a GF of 33%, while GFs ranged between 5% and 10% with all other assemblers.

4.2 Synthetic archaeal and bacterial community

Inching toward more realistic applications, a second dataset provides Illumina reads extracted from a synthetic microbial community composed of 16 archaeal species from 12 genera, as well as 48 bacterial species from 36 genera (accession SRR606249; [21]). As emphasized by the authors, the selected organisms cover a wide range of environmental conditions and adaptation strategies. In contrast to the previous simulated dataset (Section 4.1), the proportion of each species in the sample is not uniform, which results in individual genome average sequencing depth varying from 9x to 318x. The number of 16S rRNA paralogs per genome appears also highly diverse, ranging from 1 to 10 copies per genome. Altogether, this dataset represents a total amount of 106 distinct 16S rRNA sequences with pairwise sequence identities ranging from 59.64% to 99.93%.

The organisms were sequenced on Illumina HighSeq2000, providing 109 million 101bp paired-end reads with an average fragment size of 250bp. We quality cleaned the reads using Prinseq Lite [29], removed adapter sequences using Cutadapt [30], filtered out short reads (¡ 50bp), and obtained a total number of 67.6 million reads, which were analyzed with MATAM and EMIRGE. The uncleaned raw dataset was provided to REAGO, considering that the method could not handle reads with varying lengths. Finally, for SPAdes and MEGAHIT, the 16S rRNA reads were extracted from the cleaned dataset using SortMeRNA, which provided 108,560 16S rRNA reads to assemble.

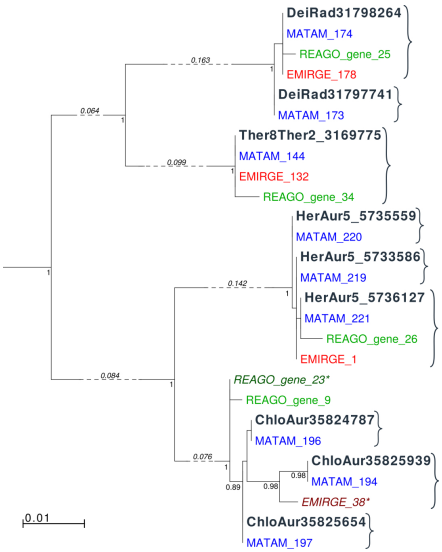


Fig. 2. Alignment of the reference sequences with the assembled contigs shows MATAM ability to differentiate between very close sequences. MATAM, EMIRGE and REAGO contigs are shown respectively in blue, red and green. In a ideal setting, each software should produce contigs that cluster closely to each reference (black) sequence. Contigs followed by a star, and drawn in a darker color, were considered as chimeric by VSEARCH.

Results are shown in Table 2. Confirming the trends observed on the simulated dataset, MATAM is able to recover the highest number of sequences together with the highest GF (83%). Most importantly, with lower ER than achieved by the other tested methods, the MATAM assembly appears highly accurate. While EMIRGE is the second best approach in terms of recovered GF, it also yields the greatest ER and Ns over all the compared tools. Moreover, a RDP classification of MATAM and EMIRGE sequences indicates that while MATAM missed one expected genus only, EMIRGE missed 4 genera out of 48.

Inspection of the MetaQuast alignments of the assemblies against the original 16S rRNAs revealed that all methods accurately assembled the genes sharing less than 90% sequence identity with their closest relatives within the sample. However, performances significantly dropped when attempting to assemble the closely related genes in the dataset. This especially concerned the paralogous 16S rRNA copies sharing around 99% sequence identity.

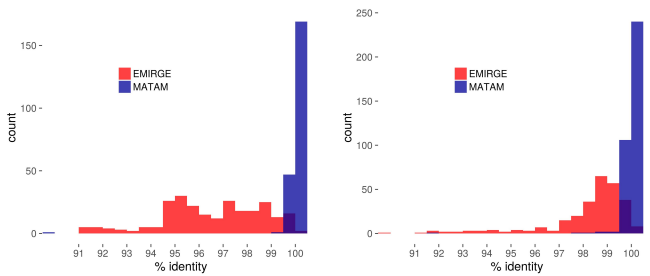
We selected sequences from a representative subset of four related species possessing one to three such paralogous copies. Those 16S rRNAs and their corresponding assembled candidate sequences were selected for a phylogenetic tree reconstruction. The obtained tree (Figure 2) demonstrates that MATAM correctly assembled all the different paralogs with nearly no error, while EMIRGE and REAGO only managed to recover one candidate sequence per species. Thus, EMIRGE and REAGO merged into a single candidate sequence the reads issued from distinct paralogs, resulting in erroneous assemblies with high ER and underestimated GF. Indeed, each of the sequences assembled with REAGO, as well as one EMIRGE sequence over four, appear to cluster at a slight distance from their respective targeted paralogs. Those distances simply account for the methods reconstruction errors. Consistently, in two cases, the candidates assembled by EMIRGE and REAGO were identified as chimeras by VSEARCH.

4.3 Human Microbiome Project

Finally, we used two metagenomic samples from the Human Microbiome Project (gut: SRS011405, and mouth: SRS016002, [22]) in order to validate MATAM on real metagenomic datasets sequenced from genuine environments. The reads were already quality cleaned and trimmed, and no additional filtering was performed. Hence, reads having different lengths, we were not able to run REAGO on these datasets. Results obtained

		Chimera (%)	#contigs	TL	#classes	#genera
SRS011405	MATAM	3.37%	218	187710	5 (4)	21 (17)
	EMIRGE	43.04%	273	393152	2 (2)	12 (8)
SRS016002	MATAM	4.92%	353	320748	13 (13)	31 (28)
	EMIRGE	46.01%	282	394087	12 (12)	25 (23)

Tab. 3. Results for the gut and mouth HMP datasets. The column *#classes* indicates the total number of taxonomic classes found with RDP from the assemblies, with the number of these classes validated with the QIIME OTUs (in parentheses). The column *#genera* gives the same information at the genus level.



(a) Human gut sample SRS011405 (b) Human mouth sample SRS016002

Fig. 3. % identity distribution of best matches against Silva 128 SSU Ref NR.

with SPAdes and MEGAHIT using the following protocol appeared highly inaccurate and therefore, they are not further commented in this work. Thus, we only present the results obtained with EMIRGE and MATAM.

For these two datasets, the exact ground truth is unknown. Thus we could not perform the same validation procedure as in the two previous examples and we had to resort to alternative strategies. First, we took advantage of the availability of OTU sequences inferred through a QIIME analysis of the V1-V3 hypervariable regions for the same biological samples (available from the SRS accession numbers). We compared the assignments obtained from assemblies, calculated with RDP, with these of amplicon OTUs (Table 3). For both samples, MATAM identified more classes and genera than EMIRGE did, and most of these taxa were validated by the amplicon OTUs. Interestingly, we observed that in the two samples, three genera were recovered both by MATAM and EMIRGE, but not by the amplicon approach: *Odoribacter*, *Peptococcus*, and *Bergeyella*. Since some species from these genera are known to be adapted to the human gut and mouth environments, it is plausible that they were missed by the amplicon approach while being accurately recovered by MATAM and EMIRGE from the metagenomic samples.

Moreover, we evaluated assembly quality by aligning MATAM and EMIRGE sequences against the complete Silva 128 SSU Ref NR database, using BLAST. The rationale for this experiment is that most of the species in these human gut and mouth samples are possibly already known, and therefore should be found in Silva. We observed that nearly all MATAM sequences matched with a known 16S rRNA in Silva with more than 99% identity, among which a majority matched with 100% identity (Figures 3a and 3b), which suggests that MATAM sequences could possibly be assigned at the species or even the strain level. On the other hand, EMIRGE sequences provided a discordant picture. In the case of the human mouth sample, most of the EMIRGE sequences obtained a match above 97% identity, but only a slight proportion of them matched with 100% identity against a known 16S rRNA (Figure 3b). The observation is even more pronounced with the human gut sample, where only 43% of the EMIRGE sequences obtained a match above 97% identity against a Silva 16S rRNA sequence (Figure 3a). Thus, conversely to MATAM, EMIRGE sequences would suggest that only a slight proportion of the human gut and mouth diversity has a known isolate registered in Silva. However, considering our previous conclusions on controlled datasets, we assume that part of this diversity inferred with EMIRGE might in fact corresponds to reconstruction artifacts.

5 Discussion

Taxonomic assignments of environmental samples is a strikingly difficult task which suffers from inherent limitations of high-throughput sequencing technologies. In this respect, we designed MATAM as an alternative to existing software helping to better understand the taxonomic structures of shotgun metagenomic samples. Our experimental results show that MATAM outperforms other available tools providing phylogenetic marker assemblies. Reconstructing full length 16S rRNAs allows to reach a higher precision of taxonomic assignments than individual read analysis or amplicon sequencing do, because the reconstructed sequences effectively contain stronger phylogenetic signal. Moreover, metagenomic shotgun sequencing is naturally immune against the primer and amplification biases attached to the amplicon sequencing technology, and therefore is more adequate to sequence unknown species.

Our approach opens up several new perspectives. Although we have focused this work on the assembly of 16S rRNA genes, MATAM was designed to deal with any marker of taxonomic interest. Indeed, there is currently an emerging trend to consider a combination of universal (single-copy) marker families, such as provided in the recently published database proGenomes [31]. Sequences from this database, or from any other customized one, could be used with MATAM to target a variety of markers, and thus provide improving taxonomic assignments. MATAM could also be used in combination with other types of sequencing data. Long read sequencing is able to produce fragments that cover large regions of the DNA molecules, up to several thousands of bases. When long reads are available, they could serve as a guide in the scaffolding step of MATAM and concomitantly, MATAM low-error contigs could be used to correct them. Finally, targeted gene capture, that allows to sequence at high depth captured DNA regions of interest from an environmental sample [32], could also prove to be an exciting application field for MATAM.

References

[1] Kenneth J. Locey and Jay T. Lennon. Scaling laws predict global microbial diversity. *Proceedings of the National Academy of Sciences*, 113(21):5970–5975, May 2016.

[2] Bo Liu, Theodore Gibbons, Mohammad Ghodsi, Todd Treangen, and Mihai Pop. Accurate and fast estimation of taxonomic profiles from metagenomic shotgun sequences. *BMC Genomics*, 12(Suppl 2):S4, July 2011.

[3] Nicola Segata, Levi Waldron, Annalisa Ballarini, Vagheesh Narasimhan, Olivier Jousson, and Curtis Huttenhower. Metagenomic microbial community profiling using unique clade-specific marker genes. *Nature Methods*, 9(8):811–814, 2012.

[4] Christian Quast, Elmar Pruesse, Pelin Yilmaz, Jan Gerken, Timmy Schweer, Pablo Yarza, Jörg Peplies, and Frank Oliver Glöckner. The SILVA ribosomal RNA gene database project: improved data processing and web-based tools. *Nucleic Acids Research*, 41(D1):D590–D596, January 2013.

[5] James R. Cole, Qiong Wang, Jordan A. Fish, Benli Chai, Donna M. McGarrell, Yanni Sun, C. Titus Brown, Andrea Porras-Alfaro, Cheryl R. Kuske, and James M. Tiedje. Ribosomal Database Project: data and tools for high throughput rRNA analysis. *Nucleic Acids Research*, 42(D1):D633–D642, January 2014.

[6] T. Z. DeSantis, P. Hugenholtz, N. Larsen, M. Rojas, E. L. Brodie, K. Keller, T. Huber, D. Dalevi, P. Hu, and G. L. Andersen. Greengenes, a Chimera-Checked 16s rRNA Gene Database and Workbench Compatible with ARB. *Applied and Environmental Microbiology*, 72(7):5069–5072, January 2006.

[7] Rachel Poretsky, Luis M. Rodriguez-R, Chengwei Luo, Despina Tsementzi, and Konstantinos T. Konstantinidis. Strengths and Limitations of 16s rRNA Gene Amplicon Sequencing in Revealing Temporal Microbial Community Dynamics. *PLoS ONE*, 9(4):e93827, April 2014.

[8] Alexander Sczyrba, Peter Hofmann, Peter Belmann, David Koslicki, Stefan Janssen, Johannes Droege, Ivan Gregor, Stephan Majda, Jessika Fiedler, Eik Dahms, Andreas Bremges, Adrian Fritz, Ruben Garrido-Oter, Tue Sparholt Jorgensen, Nicole Shapiro, Philip D. Blood, Alexey Gurevich, Yang Bai, Dmitrij Turaev, Matthew Z. DeMaere, Rayan Chikhi, Niranjan Nagarajan, Christopher Quince, Lars Hestbjerg Hansen, Soren J. Sorensen, Burton K. H. Chia, Bertrand Denis, Jeff L. Froula, Wang, and et al. Critical Assessment of Metagenome Interpretation - a benchmark of computational metagenomics software. *bioRxiv*, page 099127, January 2017.

[9] Christopher S. Miller, Brett J. Baker, Brian C. Thomas, Steven W. Singer, and Jillian F. Banfield. Emirge: reconstruction of full-length ribosomal genes from microbial community short read sequencing data. *Genome Biology*, 12(5):R44, 2011.

[10] Cheng Yuan, Jikai Lei, James Cole, and Yanni Sun. Reconstructing 16s rRNA genes in metagenomic data. *Bioinformatics*, 31(12):i35–i43, June 2015.

[11] Eric P. Nawrocki, Diana L. Kolbe, and Sean R. Eddy. Infernal 1.0: inference of RNA alignments. *Bioinformatics*, 25(10):1335–1337, May 2009.

[12] Celine Mercier, Frederic Boyer, Aurelie Bonin, and Coissac Eric. Sumatra and sumaclust: fast and exact comparison and clustering of sequences, 2013.

[13] Evguenia Kopylova, Jose A. Navas-Molina, Céline Mercier, Zhenjiang Zech Xu, Frédéric Mahé, Yan He, Hong-Wei Zhou, Torbjørn Rognes, J. Gregory Caporaso, and Rob Knight. Open-source sequence clustering methods improve the state of the art. *mSystems*, 1(1), 2016.

[14] Evguenia Kopylova, Laurent Noé, and Hélène Touzet. Sortmerna: fast and accurate filtering of ribosomal rnas in metatranscriptomic data. *Bioinformatics*, 28(24):3211–3217, 2012.

[15] Evguenia Kopylova, Laurent Noé, Pierre Pericard, Mikael Salson, and Hélène Touzet. Sortmerna 2: ribosomal rna classification for taxonomic assignation. In *Workshop on Recent Computational Advances in Metagenomics, ECCB 2014*, 2014.

[16] Jared T. Simpson and Richard Durbin. Efficient de novo assembly of large genomes using compressed data structures. *Genome Research*, 22(3):549–556, 2012.

[17] Andreas Döring, David Weese, Tobias Rausch, and Knut Reinert. Seqan an efficient, generic c++ library for sequence analysis. *BMC Bioinformatics*, 9(1):1–9, 2008.

[18] Anton Bankevich, Sergey Nurk, Dmitry Antipov, Alexey A. Gurevich, Mikhail Dvorkin, Alexander S. Kulikov, Valery M. Lesin, Sergey I. Nikolenko, Son Pham, Andrey D. Prjibelski, Alexey V. Pyshkin, Alexander V. Sirotkin, Nikolay Vyahhi, Glenn Tesler, Max A. Alekseyev, and Pavel A. Pevzner. SPAdes: A New Genome Assembly Algorithm and Its Applications to Single-Cell Sequencing. *Journal of Computational Biology*, 19(5):455–477, May 2012.

[19] 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*, 31(10):1674–1676, May 2015.

[20] Konstantinos Mavromatis, Natalia Ivanova, Kerrie Barry, Harris Shapiro, Eugene Goltsman, Alice C. McHardy, Isidore Rigoutsos, Asaf Salamov, Frank Korzeniewski, Miriam Land, Alla Lapidus, Igor Grigoriev, Paul Richardson, Philip Hugenholtz, and Nikos C. Kyrpides. Use of simulated data sets to evaluate the fidelity of metagenomic processing methods. *Nature Methods*, 4(6):495–500, June 2007.

[21] Migun Shakya, Christopher Quince, James H. Campbell, Zamin K. Yang, Christopher W. Schadt, and Mircea Podar. Comparative metagenomic and rRNA microbial diversity characterization using Archaeal and Bacterial synthetic communities. *Environmental microbiology*, 15(6):1882–1899, June 2013.

[22] The Human Microbiome Project Consortium. Structure, function and diversity of the healthy human microbiome. *Nature*, 486(7402):207–214, June 2012.

[23] Robert C. Edgar, Brian J. Haas, Jose C. Clemente, Christopher Quince, and Rob Knight. UCHIME improves sensitivity and speed of chimera detection. *Bioinformatics*, 27(16):2194–2200, August 2011.

[24] Torbjørn Rognes, Tomáš Flouri, Ben Nichols, Christopher Quince, and Frédéric Mahé. VSEARCH: a versatile open source tool for metagenomics. *PeerJ*, 4:e2584, October 2016.

[25] Alla Mikheenko, Vladislav Saveliev, and Alexey Gurevich. MetaQUAST: evaluation of metagenome assemblies. *Bioinformatics*, 32(7):1088–1090, April 2016.

[26] Qiong Wang, George M. Garrity, James M. Tiedje, and James R. Cole. Naïve Bayesian Classifier for Rapid Assignment of rRNA Sequences into the New Bacterial Taxonomy. *Applied and Environmental Microbiology*, 73(16):5261–5267, August 2007.

[27] Miguel Pignatelli and Andrés Moya. Evaluating the Fidelity of De Novo Short Read Metagenomic Assembly Using Simulated Data. *PLOS ONE*, 6(5):e19984, May 2011.

[28] Weichun Huang, Leping Li, Jason R. Myers, and Gabor T. Marth. ART: a next-generation sequencing read simulator. *Bioinformatics*, 28(4):593–594, February 2012.

[29] Robert Schmieder and Robert Edwards. Quality control and preprocessing of metagenomic datasets. *Bioinformatics*, 27(6):863–864, March 2011.

[30] Marcel Martin. Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet.journal*, 17(1):pp. 10–12, May 2011.

[31] Daniel R. Mende, Ivica Letunic, Jaime Huerta-Cepas, Simone S. Li, Kristoffer Forslund, Shinichi Sunagawa, and Peer Bork. proGenomes: a resource for consistent functional and taxonomic annotations of prokaryotic genomes. *Nucleic Acids Research*, 45(D1):D529–D534, January 2017.

[32] Cyrielle Gasc, Eric Peyretaillade, and Pierre Peyret. Sequence capture by hybridization to explore modern and ancient genomic diversity in model and nonmodel organisms. *Nucleic Acids Research*, 44(10):4504–4518, June 2016.

FEELnc: an alignment-free tool for long non-coding RNAs annotation

Valentin WUCHER, Fabrice LEGEAI[2,3], Benoît HÉDAN[1], Guillaume RIZK[3], Lætitia LAGOUTTE[1], Édouard CADIEU[1], Audrey DAVID[2], Nadine BOTHEREL[1], Céline LE BÉGUEC[1], Catherine ANDRÉ[1], Christophe HITTE[1] ,Thomas DERRIEN[1]

[1]IGDR, CNRS, UMR6290, University Rennes1, Rennes, Cedex 35043, France
[2]IGEPP, BIPAA, INRA, Campus Beaulieu, Le Rheu 35653, France
[3]INRIA, Genscale, Campus Beaulieu, Rennes 35042, France

Corresponding Author: tderrien@univ-rennes1.fr

Paper Reference: **Wucher *et al.* (2017) FEELnc: a tool for long non-coding RNA annotation and its application to the dog transcriptome. *Nucleic Acids Research*. https://doi.org/10.1093/nar/gkw1306**

Abstract *Whole transcriptome sequencing (RNA-seq) has become a standard for cataloguing and monitoring RNA populations. One of the main bottlenecks consists in correctly identifying the different classes of RNAs among the plethora of reconstructed transcripts, particularly those that will be translated (mRNAs) from the class of long non-coding RNAs (lncRNAs). Here, we present FEELnc (FlExible Extraction of LncRNAs), an alignment-free program that accurately annotates lncRNAs based on a Random Forest model trained with general features such as multi k-mer frequencies and relaxed open reading frames. Benchmarking versus five state-of-the-art tools shows that FEELnc achieves similar or better classification performance on gold standard annotation datasets. Importantly, the program provides specific modules that enable the user to fine-tune classification accuracy, to formalize the annotation of lncRNAs and to identify lncRNAs even in the absence of an lncRNA training set. FEELnc moves beyond conventional coding potential classifiers by providing a standardized and complete solution for annotating lncRNAs. FEELnc is freely available at https://github.com/tderrien/FEELnc.*

Keywords Long non-coding RNAs, Machine learning, Non-model species, Classification.

1 FEELnc description

FEELnc is an all-in-one solution from the filtering of non-lncRNA-like transcript models, to the computation of a new coding potential score and the automation of the definition of lncRNA classes. Based on a relaxed definition of ORFs and a very fast analysis of *k*-mer frequencies, the program implements an alignment-free strategy using Random Forests [1] to classify lncRNAs and mRNAs. Particularly, we developed FEELnc to be used on organisms for which no set of lncRNAs is available by deriving species-specific lncRNA models from mRNA sequences and automatically computing the coding potential score cut-off that maximizes classification performances.

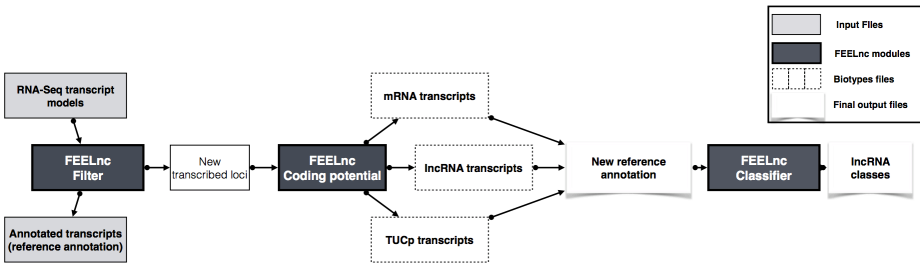


Fig 1. FEELnc workflow and its 3 modules (from left to right): Filter, Coding_potential, Classifier.

The FEELnc workflow (Fig. 1) starts with a module, the filter, which aims at identifying non-lncRNA transcripts from the reconstructed transcript models given by transcriptome reconstruction tools [2]. The

second FEELnc module aims at computing a coding potential score given the assembled sequences and thus discriminates lncRNAs/mRNAs. Finally, the third FEELnc module, the classifier, employs a sliding window strategy around each lncRNAs to report all the reference biotypes/transcripts located within the window.

2 FEELnc benchmarks

In order to compare the performance of FEELnc with state-of-the art methods, we benchmarked tools based on human and mouse gold-standard GENCODE annotations [3]. In summary, FEELnc shows similar or better performance compared to five alignment-based and alignment-free tools (Table 1).

Programs	Alignment	Sensitivity	Specificity	Precision	Accuracy	MCC
FEELnc	no	**0.923**	0.915	0.916	**0.919**	**0.838**
CPAT	no	0.899	0.924	0.922	0.912	0.823
CNCI	no	0.829	0.979	0.975	0.904	0.817
PLEK	no	0.732	**0.985**	**0.981**	0.858	0.741
PhyloCSF	yes	0.906	0.802	0.820	0.854	0.712
CPC	yes	0.699	0.739	0.728	0.719	0.438

Table 1. Tools performance on the human data sets (similar results were observed in mouse). Bold values are the highest for each metrics. Programs are ranked by MCC (Matthews Correlation Coefficient).

3 Training FEELnc without known lncRNAs

One issue when using machine-learning approach is the requirement of both a positive and a negative set (here mRNA and lncRNA) to train the model. While the former is often available for most organisms, the latter is usually not, especially for many organisms [4]. To simulate lncRNAs, we also assessed two strategies called "shuffle" and "cross-species". The shuffle strategy is based on the paradigm that lncRNAs are derived from 'debris' of protein-coding genes [5]. To this end, we shuffled mRNA sequences from a reference annotation using the Ushuffle program [6], while preserving a given *k*-mer frequency of the input sequences. The cross-species strategy makes use of lncRNA sets annotated in other species to extract non-coding predictors and train the Random Forest model.

As expected, we showed that FEELnc performance for the cross-species strategy decreases with the time of speciation between the targeted species and the species providing lncRNAs (Sp_{rho}= -0.85, $p_{val}<10^{-4}$) and the shuffle strategy is a valuable alternative when no lncRNAs are annotated in closely related species.

4 Conclusion

We present FEELnc, a new program to annotate lncRNAs based on RNA-Seq assembled transcripts. In addition to providing good performance metrics, FEELnc allows to be self-trained by own users' datasets and also can be used for non-model organisms for which no set of lncRNAs is annotated.

References

[1] Breiman L. Random Forests. Machine Learning. *Kluwer Academic Publishers*; 2001;45: 5–32. doi:10.1023/A:1010933404324

[2] Trapnell C, Williams BA, Pertea G, Mortazavi A, Kwan G, van Baren MJ, et al. Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nature Biotechnology*. 2010;28: 511–515. doi:10.1038/nbt.1621

[3] Harrow J, Frankish A, Gonzalez JM, Tapanari E, Diekhans M, Kokocinski F, et al. GENCODE: The reference human genome annotation for The ENCODE Project. *Genome Res*. 2012;22: 1760–1774. doi:10.1101/gr.135350.111

[4] Tagu D, Colbourne JK, Nègre NN. Genomic data integration for ecological and evolutionary traits in non-model organisms. *BMC Genomics*. 2014;15: 1–16. doi:10.1186/1471-2164-15-490

[5] Duret L, Chureau C, Samain S, Weissenbach J, Avner P. The Xist RNA gene evolved in eutherians by pseudogenization of a protein-coding gene. *Science*. 2006;312: 1653–1655. doi:10.1126/science.1126316

[6] Jiang M, Anderson J, Gillespie J, Mayne M. uShuffle: A useful tool for shuffling biological sequences while preserving the k-let counts. *BMC Bioinformatics*; 2008;9: 192–11. doi:10.1186/1471-2105-9-192

Analyse intégrative des ARN longs non-codants (lncRNAs) du génome canin

Céline LE BÉGUEC[1], Valentin WUCHER[1,2], Lætitia LAGOUTTE[1], Edouard CADIEU[1], Benoît HÉDAN[1], Catherine ANDRÉ[1], Christophe HITTE[1], Thomas DERRIEN[1]

[1] Institut de Génétique et Développement de Rennes, CNRS UMR6290, Université Rennes1, 35000 Rennes, France. [2] Adresse actuelle : Centre for Genomic Regulation (CRG), The Barcelona Institute of Science and Technology, Dr. Aiguader 88, Barcelona 08003, Spain.

Auteur pour correspondance : Thomas.derrien@univ-rennes1.fr

Résumé : *Les ARN longs non codants (lncRNAs) constituent une famille d'ARN hétérogènes qui jouent un rôle majeur dans de nombreux processus biologiques. Nous avons développé récemment un outil d'annotation des lncRNAs appelé FEELnc, basé sur une approche sans alignement, et utilisant une stratégie 'Random Forest' entraînée sur les fréquences de multiples k-mer. Dans le cadre du consortium européen LUPA, FEELnc a permis d'identifier plusieurs milliers de nouveaux lncRNAs du génome du chien mais leur annotation fonctionnelle demeure dans son ensemble mal caractérisée. Dans ce travail, nous avons produit et analysé les profils d'expression des lncRNAs canins (n >10000) sur 26 RNA-seq représentant une grande diversité de tissus. Nous présentons une caractérisation fonctionnelle des lncRNAs canins portant sur l'identification de leur signature d'expression spécifique par tissu, l'analyse de leurs rôles potentiels comme régulateurs transcriptionnels, leurs niveaux de conservation et leurs contenus en éléments transposables. Nous avons identifié 4600 lncRNAs présentant un patron d'expression tissu-spécifique, avec près de 63 lncRNAs exprimés spécifiquement par tissu, suggérant un rôle essentiel dans la genèse et le maintien de l'intégrité des tissus. Nous avons construit un réseau de co-expression de l'ensemble des paires lncRNA:mRNA pour analyser des relations de cis-régulation. L'analyse statistique des paires a permis de déterminer des corrélations significatives (p.adjust BH <0,01) qui suggèrent un rôle régulateur pour plus de 900 lncRNAs. L'analyse des mRNAs cibles identifiés par les corrélations, basée sur les termes GO, a permis d'identifier des annotations fonctionnelles significativement enrichies ($p < 1^{e-4}$, FDR <0.05) correspondant aux processus de développement tels que le développement d'organes sensoriels et la croissance cellulaire. L'analyse de la co-occurence des éléments transposables et des lncRNAs canins a montré que 84% des lncRNAs contiennent des SINEs spécifiques aux canidés. Ces co-occurences SINE:lncRNA suggèrent l'importance des éléments transposables dans l'expression tissu-spécifique des lncRNAs et dans la biogénèse des lncRNAs identifiés chez le chien.*

Mots clés : lncRNAs, transcriptome, co-expression, éléments transposables, chien.

1. Introduction

Avec l'avancée des nouvelles technologies de séquençage haut débit, les analyses du transcriptome (RNA-seq) permettent d'identifier l'ensemble des classes d'ARN dont la classe des ARN longs non codants (lncRNAs) [1,2]. Le transcriptome correspond ainsi à l'ensemble des molécules d'ARN transcrites, avec ou sans capacité de coder des protéines, pour un temps et une condition ou tissu donnés [3]. Arbitrairement définis selon un critère de taille (généralement plus de 200 nucléotides), les lncRNAs possèdent des caractéristiques similaires à celles des ARN codant pour des protéines (mRNAs), c'est-à-dire qu'ils peuvent être épissés, et posséder (ou non) une queue de polyadénylation, mais ils se différencient par une absence de cadre de lecture ouvert fonctionnel. A la suite du séquençage de l'ensemble des transcrits ARN, l'annotation et la classification des différents ARN consiste à reconstruire les modèles de transcrits, à partir desquels il est crucial de définir les classes fonctionnelles des nouveaux transcrits identifiés. Les lncRNAs annotés peuvent

être localisés soit dans des régions intergéniques (lincRNAs), soit chevauchant et anti-sens des mRNAs (AS-lncRNAs) et représentent de bons candidats pour assurer un rôle de régulateur notamment des mRNAs proximaux.

Le modèle canin a émergé récemment comme une ressource nouvelle pour étudier la base génétique des traits complexes, incluant la morphologie, la physiologie et le comportement [4]. De multiples ressources génétiques et une séquence génomique de haute qualité positionnent désormais le chien comme une espèce modèle importante pour comprendre l'évolution du génome, la génétique de la population canine et les gènes sous-jacents aux traits phénotypiques complexes. Cependant, si les ressources génomiques nouvellement développées ont élargi notre compréhension du génome canin, l'annotation exhaustive des éléments fonctionnels tels que les ARN régulateurs de l'expression des gènes demeure nécessaire pour faciliter l'identification des relations génotype-phénotype [5]. Une des particularités du génome du chien réside dans la présence d'une famille d'éléments transposables spécifiques (SINEC_Cf) [6]. Ces rétrotransposons dégénérés (éléments auto-réplicants utilisant un ARN intermédiaire) sont dérivés d'un ARNt-Lys et ont été impliqués dans de nombreuses maladies génétiques [7,8] ou différences phénotypiques entre races de chiens [9,10]. Chez l'homme et la souris, les éléments transposables sont fréquemment retrouvés dans les lncRNAs [11,12] et de nombreuses études soulignent l'importance des éléments transposables dans la régulation de l'expression des gènes (voir pour revue récente [13]).

Nous avons récemment développé un outil appelé FEELnc qui analyse l'ensemble des transcrits et permet d'extraire, d'annoter et de classifier les ARN longs non codants [14]. Nous avons utilisé FEELnc dans le cadre du programme européen LUPA (http://eurolupa.org/), un consortium dédié au modèle canin, pour étendre l'annotation du répertoire des lncRNAs à partir de 26 RNA-seq de tissus distincts canins. Nous présentons dans cette étude une vaste analyse d'expression des >10000 lncRNAs chez le chien, leur corrélation d'expression avec les ARN messagers, et l'analyse de leur conservation et de leur contenu en éléments transposables. Nos résultats précisent la classification et les caractéristiques d'expression des lncRNAs au regard des mRNAs, la nature spécifique de leur expression et identifient un sous-ensemble de lncRNAs conservés entre l'homme et le chien.

2. Analyse du patron d'expression des lncRNAs

Nous avons produit et utilisé les données de l'annotation canine 'canFam3.1-plus' comportant 10444 gènes lncRNAs et 21810 gènes mRNAs [14]. Selon le protocole bioinformatique décrit dans Djebali et al. [15], nous avons aligné les lectures de 26 échantillons RNA-seq (>30 millions/RNA-seq), séquencés par HiSeq2000 à partir de banque cDNAs orientés, sur le génome de référence canin et à l'aide de l'annotation canFam3.1-plus avec le programme STAR (STAR_2.5.0a) [16] et avons déterminé l'expression des gènes et isoformes lncRNAs et mRNAs pour les 26 RNA-seq avec le programme RSEM (RSEM-1.2.25) [17]. L'analyse des 26 échantillons canins, représentant une grande diversité de tissus, a mis en évidence que 75% (7764/10444) des gènes lncRNAs étaient exprimés dans au moins un tissu (Transcripts Per Million i.e TPM >1). En comparant les distributions des niveaux d'expression entre gènes lncRNAs et les mRNAs, nous observons, comme chez l'homme [2] ou la souris [18] que les lncRNAs ont une plus faible expression (au moins 20 fois inférieure) pour tous les tissus analysés, à l'exception notable du tissu testiculaire ou la différence de moyennes d'expressions entre lncRNAs et mRNAs est moins importante (6 fois) (Fig. 1A).

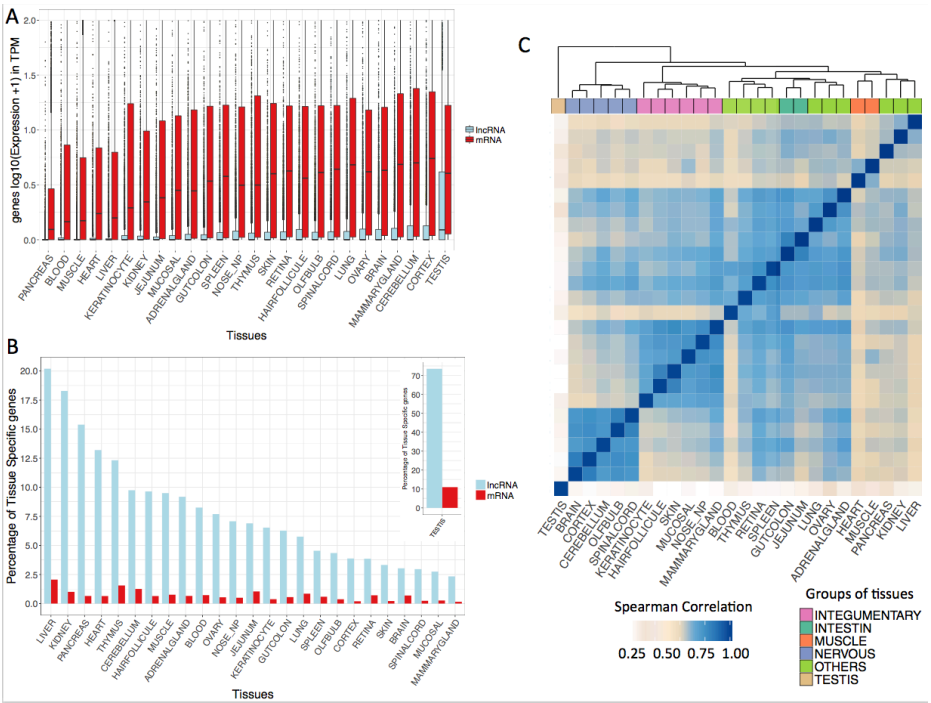


Fig 1. (A) Analyse comparée des niveaux de transcription en log (TPM+1) entre gènes mRNAs (rouge foncé) et lncRNAs (bleu clair) dans 26 tissus. **(B)** Proportions de gènes lncRNAs (bleu clair) et mRNAs (rouge foncé) tissu-spécifique (le tissu testiculaire est représenté dans l'encadré étant donné la forte proportion de gènes exprimés spécifiquement dans ce tissu) **(C)** Clustering hiérarchique des 26 tissus basé sur les corrélations de Spearman mesurées à partir des données d'expression lncRNAs.

Nous avons déterminé le niveau de tissu-spécificité par le calcul d'un score de spécificité tissulaire, "*tau*" [19], pour chaque gène lncRNAs et mRNAs (dont le niveau d'expression TPM >1) et avons ensuite mené une étude comparative de la spécificité tissulaire pour les 26 tissus. Ce score de spécificité, compris entre 0 et 1 (*tau* = 1 pour les gènes très spécifiques) est défini selon l'équation suivante :

$$\tau = \frac{\sum_{i=1}^{n}(1-\hat{x}_i)}{n-1}, \hat{x}_i = \frac{x_i}{\max_{1\leq i\leq n}(x_i)}$$

avec n correspondant au nombre de tissus analysés et x_i, l'expression du gène dans le tissu i. Dans une récente étude comparative, ce score a été classé parmi les plus robustes pour annoter les gènes tissu-spécifiques (TS) [20]. En imposant un filtre strict tau ≥ 0,95 (qui correspond en moyenne à un ratio $\max(x_i)$ / $\max_{2nd}(x_i)$ supérieur à 4), un total de 4600 lncRNAs tissu-spécifiques et 3600 mRNAs tissu-spécifiques ont été identifiés. Nous observons en moyenne 63 lncRNAs exprimés spécifiquement par tissu, en excluant le tissu testiculaire. Ces sous-ensembles de profils d'expression, majoritairement lncRNAs, révèlent de véritables signatures associées à chaque tissu qui suggèrent des fonctions spécifiques dans les

processus physiologiques des tissus (Fig. 1B). Comme chez l'homme [1,2], le tissu testiculaire est particulièrement enrichi en gènes tissus-spécifiques (~3000 lncRNAs), ce qui met en évidence la singularité de ce tissu, probablement dû à l'état plus relâché de sa chromatine et à la présence de nombreux types cellulaires [21]. Par extension, l'utilisation des lncRNAs tissu-spécifique dans le cadre d'une analyse de la variabilité d'expression ou du différentiel d'expression pourra permettre d'investiguer leur implication fonctionnelle liée à l'état pathologique ou à l'environnement.

À partir des corrélations de Spearman mesurées sur les données d'expression lncRNAs, nous avons construit une matrice de distance et réalisé un clustering hiérarchique (méthode d'agrégation de ward) (Fig. 1C), qui met à nouveau en évidence la spécificité du tissu testiculaire par rapport aux 25 autres tissus. De plus, le clustering permet aussi de visualiser le regroupement cohérent des tissus ayant des origines histologiques communes et/ou des proximités fonctionnelles comme les tissus tégumentaires, musculaires ou nerveux (Fig. 1C). Par exemple, nous identifions le lncRNA RLOC_00033166 exprimé dans les 5 tissus nerveux (TPM >2) et non transcrit dans tous les autres tissus. Ce lncRNA est par ailleurs localisé en anti-sens du gène *NRG3* (Neuregulin 3) qui est impliqué dans la différentiation cellulaire des neuroblastes et représente donc un candidat potentiel à sa régulation.

3. Construction et analyse du réseau de co-expression

La majorité des lncRNAs qui ont pu être caractérisés chez d'autres espèces possèdent une fonction de régulation de l'expression d'un gène en coordonnant des processus épigénétiques, transcriptionnels ou post-transcriptionnels [22]. Des analyses statistiques de corrélations positives ou négatives des valeurs d'expression entre mRNAs et lncRNAs peuvent permettre de mettre en valeur des lncRNAs régulant l'expression de gènes codant pour des protéines sachant que la plupart des interactions lncRNA:mRNA validées fonctionnellement concernent des gènes qui se localisent dans une certaine proximité génomique (généralement <1 Mb) [23]. Par conséquent, la caractérisation positionnelle génomique [24] des lncRNAs vis-à-vis des mRNAs est une étape initiale et essentielle pour analyser des relations de *cis*-régulation potentielles. En utilisant le module de classification du programme FEELnc pour identifier l'ensemble des paires lncRNA:mRNA, nous avons pu identifier 9615 interactions classées selon 2 types (i) AS-lncRNA:mRNA défini par les lncRNAs chevauchant des mRNAs transcrit en anti-sens (n = 4531) et (ii) lincRNAs:mRNA défini par les lncRNAs intergéniques localisés à moins d'1 Mb d'un mRNA (n = 5083). En utilisant les données d'expression issues des 26 RNA-seq, nous avons ensuite appliqué le concept de guilt-by-association (coupable par association), principe qui repose sur l'observation que les transcrits co-exprimés sont plus susceptibles d'être co-régulés, de partager des fonctions similaires ou de participer à des processus biologiques similaires [25]. Les analyses statistiques de corrélations des 9615 données de co-expression, ont permis de déterminer des corrélations significatives (p.adjust BH <0,01) pour 492 paires de type lincRNA:mRNA et pour 411 paires AS-lncRNA:mRNA.

Pour attribuer des fonctions potentielles aux lncRNAs, nous avons mené une étude d'enrichissement des termes GO (Biological Process) des mRNAs avec lesquels ils sont co-exprimés. Un ensemble de 12 termes GO ont été retrouvés significativement enrichis ($p < 1^{e-4}$; FDR <0,05) et correspondent aux processus de développement tels que 'sensory organ development' (GO:0007423, 28 gènes) et 'cell growth' (GO:0016049, 25 gènes). Ces résultats permettent de proposer une assignation fonctionnelle aux lncRNAs, bien que les prédictions fondées sur ces enrichissements sont dépendants des annotations sous-jacentes.

4. Analyse de la conservation des lncRNAs canins avec l'homme

Pour identifier des lncRNAs orthologues entre l'homme et le chien qui vont suggérer des fonctions potentiellement conservées entre espèces, nous avons utilisé la base de données Compara d'EnsEMBL [26] qui recense les régions synténiques orthologues via un alignement multiple complet de génomes de plusieurs espèces. Cet alignement, qui ne prend pas en compte le sens de transcription, implique de restreindre l'analyse aux lincRNAs (n = 5651) et ne pas considérer les AS-lncRNAs. L'analyse a donc consisté à cartographier les positions des lincRNAs canins sur le génome humain via Compara, et a permis d'identifier

727 (12,86%) lincRNAs possédant un orthologue humain. Par comparaison, l'analyse pour les mRNAs a déterminé que 66,02% des mRNAs ont un orthologue humain selon la même méthodologie.

L'analyse des relations d'orthologie nous a permis d'étudier la distribution des scores de spécificité tissulaire (*tau*) en fonction de leur conservation au cours de l'évolution. Nous avons ainsi observé que les gènes sans relation d'orthologie sont significativement plus tissu-spécifique que les gènes conservés (test wilcoxon ; $p < 2{,}2^{e\text{-}16}$). Ces résultats montrent que les gènes possédant une expression localisée dans un type cellulaire ou un tissu sont également des gènes moins conservés au cours de l'évolution [27]. Ces observations suggèrent que ces gènes, en grande majorité les lincRNAs évoluent rapidement en terme de séquence, et ainsi se différencient pour devenir des acteurs spécifiques du développement, de la signalisation et de la régulation des principaux types et fonctions cellulaires.

5. LncRNAs et éléments transposables canins

Chez l'homme, 80% des lncRNAs chevauchent au moins un élément transposable (TE) et près de 40% des séquences lncRNAs sont dérivées de TEs [11,12] suggérant un rôle essentiel des TEs dans la genèse et l'évolution des lncRNAs [28]. Une des particularités du génome du chien réside dans la présence d'une famille spécifique de TEs, les SINEC_Cf, qui ont été montrés comme liés à la diversité phénotypique observée entre les races de chiens [6,9,10] et à l'origine de maladies [7,8].

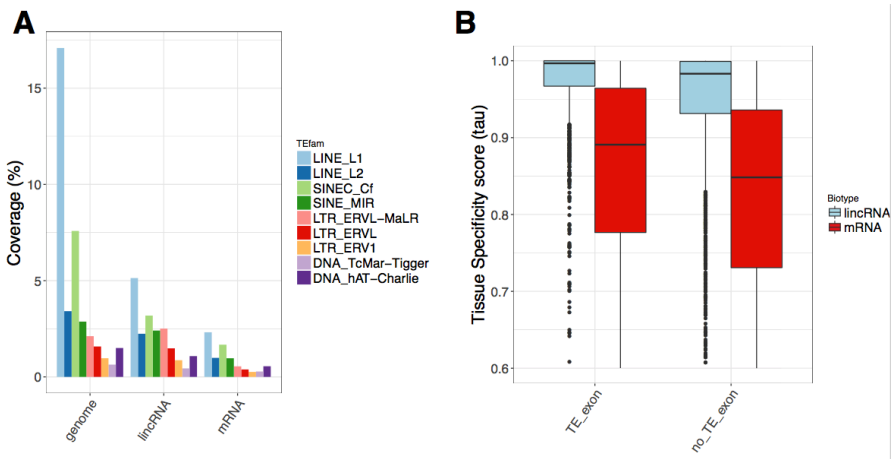


Fig 2. Analyse des éléments transposables (TEs). **(A)** Couverture du génome canin, des lincRNAs et mRNAs pour différentes classes et familles de TEs. **(B)** Distribution du score de spécificité tissulaire en fonction de la présence ou non de TEs dans les lincRNAs (bleu clair) et mRNAs (rouge foncé).

Nous avons croisé l'annotation RepeatMasker [29] de 4 classes (et leurs familles principales) d'éléments transposables canins (DNA transposons, LTRs, LINEs et SINEs) et montrons que ~84% (8793/10444) des lncRNAs canins contiennent au moins un TE et que ~20% des séquences exoniques des lncRNAs sont composées exclusivement de TEs (Fig. 2A). Cette proportion en TEs est inférieure à celle observée dans la totalité du génome du chien (37,7%) mais 2,5 fois plus importante que celle des séquences mRNAs (7,9%) mettant en évidence la forte prévalence des TEs dans les séquences lncRNAs. Comparé au génome, la plupart des familles de TEs sont sous-représentées dans les lincRNAs à l'exception des rétrovirus ERVL-MaLR (Fig. 2A) qui sont aussi significativement enrichis dans les lncRNAs humains [12]. L'analyse des données d'expression sur les 26 tissus de cette étude montrent que les gènes contenant des TEs sont transcrits de manière plus tissu-spécifique (Fig. 2B) que ceux ne contenant aucun TEs. Ces observations qui montrent la corrélation entre l'insertion de TEs et une expression tissu-spécifique, suggèrent que les

modifications rapides au niveau de la séquence des gènes par insertion de TEs vont concourir à une spécialisation fonctionnelle notamment des lncRNAs.

6. Conclusions

Nous reportons dans cette étude l'analyse d'expression de plus de 10000 lncRNAs canins dans 26 tissus distincts. Les résultats montrent que les lncRNAs canins, comme pour les autres espèces, sont exprimés plus faiblement que les mRNAs mais de manière spécifique à un tissu. Ainsi, nous avons identifié ici les signatures transcriptionnelles spécifiques aux 26 tissus analysés. La forte spécificité tissulaire des lncRNAs peut constituer un facteur limitant pour annoter le répertoire des lncRNAs d'une espèce puisque leur annotation va ainsi dépendre de la disponibilité de nombreux types cellulaires, de l'analyse de nombreux tissus et de multiples conditions de temporalité, pour définir de manière exhaustive le catalogue complet des lncRNAs et de leur isoformes.

L'étude des réseaux de co-expression révèle plus de 900 corrélations d'expression avec les gènes codant pour des protéines proches et met en évidence de possibles processus de *cis*-régulation géniques [30]. Par une approche de génomique comparative, nous identifions plus de 700 lincRNAs avec un orthologue humain et l'analyse du contenu en éléments transposables montre que ~84% des lncRNAs canins contiennent au moins un TE. Nous montrons que ces caractéristiques de conservation et de contenu en TE corrèlent avec une expression tissu-spécifique. Ces observations suggèrent que les lncRNAs canins évoluent rapidement en terme de séquence (insertion de SINEC_Cf) et ainsi se différencient pour devenir des éléments fonctionnels spécifiques du développement et de la régulation de fonctions cellulaires spécialisées.

Cette étude indique que les lncRNAs appartiennent à de multiples classes fonctionnelles et suggère des fonctions potentielles aux lncRNAs, dans des processus fondamentaux tels que la spermatogenèse et le développement, mais aussi dans des mécanismes plus spécifiques tels que le développement d'organes sensoriels et la croissance cellulaire.

Remerciements

Nous remercions l'ensemble des collègues de l'équipe Génétique du Chien de l'IGDR et Sarah Djebali de l'INRA GenPhyse de Toulouse. Nous remercions Cani-DNA (ANR-11-INBS-0003) pour les échantillons canins, le consortium LUPA (http://eurolupa.org) et le BROAD Institute pour la partie séquençage.

Références

[1] M. N. Cabili, C. Trapnell, L. Goff, M. Koziol, B. Tazon-Vega, A. Regev, and J. L. Rinn, "Integrative annotation of human large intergenic noncoding RNAs reveals global properties and specific subclasses.," *Genes & Development*, vol. 25, no. 18, pp. 1915–1927, Sep. 2011.

[2] T. Derrien, R. Johnson, G. Bussotti, A. Tanzer, S. Djebali, H. Tilgner, G. Guernec, D. Martin, A. Merkel, D. G. Knowles, J. Lagarde, L. Veeravalli, X. Ruan, Y. Ruan, T. Lassmann, P. Carninci, J. B. Brown, L. Lipovich, J. M. Gonzalez, M. Thomas, C. A. Davis, R. Shiekhattar, T. R. Gingeras, T. J. Hubbard, C. Notredame, J. Harrow, and R. Guigo, "The GENCODE v7 catalog of human long noncoding RNAs: analysis of their gene structure, evolution, and expression.," *Genome Res*, vol. 22, no. 9, pp. 1775–1789, Sep. 2012.

[3] Z. Wang, M. Gerstein, and M. Snyder, "RNA-Seq: a revolutionary tool for transcriptomics.," *Nature Publishing Group*, vol. 10, no. 1, pp. 57–63, Jan. 2009.

[4] E. K. Karlsson and K. Lindblad-Toh, "Leader of the pack: gene mapping in dogs and other model organisms," *Nature Publishing Group*, vol. 9, no. 9, pp. 713–725, Sep. 2008.

[5] L. Andersson, A. L. Archibald, C. D. Bottema, R. Brauning, S. C. Burgess, D. W. Burt, E. Casas, H. H. Cheng, L. Clarke, C. Couldrey, B. P. Dalrymple, C. G. Elsik, S. Foissac, E. Giuffra, M. A. Groenen, B. J. Hayes, L. S. Huang, H. Khatib, J. W. Kijas, H. Kim, J. K. Lunney, F. M. McCarthy, J. C. McEwan, S. Moore, B. Nanduri, C. Notredame, Y. Palti, G. S. Plastow, J. M. Reecy, G. A. Rohrer, E. Sarropoulou, C. J. Schmidt, J. Silverstein, R. L. Tellam, M. Tixier-Boichard, G. Tosser-Klopp, C. K. Tuggle, J. Vilkki, S. N. White, S. Zhao, H. Zhou, FAANG Consortium, "Coordinated international action to accelerate genome-to-phenome

with FAANG, the Functional Annotation of Animal Genomes project.," *Genome Biol*, vol. 16, no. 1, p. 57, Mar. 2015.

[6] W. Wang, "Short interspersed elements (SINEs) are a major source of canine genomic diversity," *Genome Res*, vol. 15, no. 12, pp. 1798–1808, Dec. 2005.

[7] L. Lin, J. Faraco, R. Li, H. Kadotani, W. Rogers, X. Lin, X. Qiu, P. J. de Jong, S. Nishino, and E. Mignot, "The sleep disorder canine narcolepsy is caused by a mutation in the hypocretin (orexin) receptor 2 gene," vol. 98, no. 3, pp. 365–376, Aug. 1999.

[8] M. Pelé, L. Tiret, J.-L. Kessler, S. Blot, and J.-J. Panthier, "SINE exonic insertion in the PTPLA gene leads to multiple splicing defects and segregates with the autosomal recessive centronuclear myopathy in dogs.," *Human Molecular Genetics*, vol. 14, no. 11, pp. 1417–1427, Jun. 2005.

[9] B. Hédan, S. Corre, C. Hitte, S. Dréano, T. Vilboux, T. Derrien, B. Denis, F. Galibert, M.-D. Galibert, and C. André, "Coat colour in dogs: identification of the merle locus in the Australian shepherd breed.," *BMC Vet. Res.*, vol. 2, p. 9, 2006.

[10] H. G. Parker, B. M. Vonholdt, P. Quignon, E. H. Margulies, S. Shao, D. S. Mosher, T. C. Spady, A. Elkahloun, M. Cargill, P. G. Jones, C. L. Maslen, G. M. Acland, N. B. Sutter, K. Kuroki, C. D. Bustamante, R. K. Wayne, and E. A. Ostrander, "An expressed fgf4 retrogene is associated with breed-defining chondrodysplasia in domestic dogs.," *Science*, vol. 325, no. 5943, pp. 995–998, Aug. 2009.

[11] A. Kapusta, Z. Kronenberg, V. J. Lynch, X. Zhuo, L. Ramsay, G. Bourque, M. Yandell, and C. Feschotte, "Transposable Elements Are Major Contributors to the Origin, Diversification, and Regulation of Vertebrate Long Noncoding RNAs," *PLoS Genet*, vol. 9, no. 4, p. e1003470, Apr. 2013.

[12] D. Kelley and J. Rinn, "Transposable elements reveal a stem cell-specific class of long noncoding RNAs.," *Genome Biol*, vol. 13, no. 11, p. R107, Nov. 2012.

[13] E. B. Chuong, N. C. Elde, and C. Feschotte, "Regulatory activities of transposable elements: from conflicts to benefits," *Nature Publishing Group*, vol. 18, no. 2, pp. 71–86, Feb. 2017.

[14] V. Wucher, F. Legeai, B. Hédan, G. Rizk, L. Lagoutte, T. Leeb, V. Jagannathan, E. Cadieu, A. David, H. Lohi, S. Cirera, M. Fredholm, N. Botherel, P. A. J. Leegwater, C. Le Beguec, H. Fieten, J. Johnson, J. Alföldi, C. André, K. Lindblad-Toh, C. Hitte, and T. Derrien, "FEELnc: a tool for long non-coding RNA annotation and its application to the dog transcriptome.," *Nucleic Acids Res*, p. gkw1306, Jan. 2017.

[15] S. Djebali, V. Wucher, S. Foissac, C. Hitte, E. Corre, and T. Derrien, "Bioinformatics Pipeline for Transcriptome Sequencing Analysis.," *Methods Mol. Biol.*, vol. 1468, pp. 201–219, 2017.

[16] A. Dobin, C. A. Davis, F. Schlesinger, J. Drenkow, C. Zaleski, S. Jha, P. Batut, M. Chaisson, and T. R. Gingeras, "STAR: ultrafast universal RNA-seq aligner.," *Bioinformatics*, vol. 29, no. 1, pp. 15–21, Jan. 2013.

[17] B. Li and C. N. Dewey, "RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome.," *BMC Bioinformatics*, vol. 12, no. 1, p. 323, Aug. 2011.

[18] D. D. Pervouchine, S. Djebali, A. Breschi, C. A. Davis, P. P. Barja, A. Dobin, A. Tanzer, J. Lagarde, C. Zaleski, L.-H. See, M. Fastuca, J. Drenkow, H. Wang, B. Pei, S. Balasubramanian, J. Monlong, A. Harmanci, M. Gerstein, M. A. Beer, C. Notredame, R. G. oacute, and T. R. Gingeras, "Enhanced transcriptome maps from multiple mouse tissues reveal evolutionary constraint in gene expression," *Nat Commun*, vol. 6, pp. 1–11, Jan. 2015.

[19] I. Yanai, H. Benjamin, M. Shmoish, V. Chalifa-Caspi, M. Shklar, R. Ophir, A. Bar-Even, S. Horn-Saban, M. Safran, E. Domany, D. Lancet, and O. Shmueli, "Genome-wide midrange transcription profiles reveal expression level relationships in human tissue specification.," *Bioinformatics*, vol. 21, no. 5, pp. 650–659, Mar. 2005.

[20] N. Kryuchkova-Mostacci and M. Robinson-Rechavi, "A benchmark of gene expression tissue-specificity metrics.," *Briefings in Bioinformatics*, Feb. 2016.

[21] F. Chalmel and A. D. Rolland, "Linking transcriptomics and proteomics in spermatogenesis.," *Reproduction*, vol. 150, no. 5, pp. R149–57, Nov. 2015.

[22] A. C. Mallory and A. Shkumatava, "LncRNAs in vertebrates: Advances and challenges.," *Biochimie*, Mar. 2015.

[23] A. R. Bassett, A. Akhtar, D. P. Barlow, A. P. Bird, N. Brockdorff, D. Duboule, A. Ephrussi, A. C. Ferguson-Smith, T. R. Gingeras, W. Haerty, D. R. Higgs, E. A. Miska, and C. P. Ponting, "Considerations when investigating lncRNA function in vivo.," *Elife*, vol. 3, p. e03058, Aug. 2014.

[24] I. Ulitsky, "Evolution to the rescue: using comparative genomics to understand long non-coding RNAs.," *Nature Publishing Group*, vol. 17, no. 10, pp. 601–614, Oct. 2016.

[25] M. Guttman, I. Amit, M. Garber, C. French, M. F. Lin, D. Feldser, M. Huarte, O. Zuk, B. W. Carey, J. P. Cassady, M. N. Cabili, R. Jaenisch, T. S. Mikkelsen, T. Jacks, N. Hacohen, B. E. Bernstein, M. Kellis, A. Regev, J. L. Rinn, and E. S. Lander, "Chromatin signature reveals over a thousand highly conserved large non-coding RNAs in mammals.," *Nature*, vol. 458, no. 7235, pp. 223–227, Mar. 2009.

[26] J. Herrero, M. Muffato, K. Beal, S. Fitzgerald, L. Gordon, M. Pignatelli, A. J. Vilella, S. M. J. Searle, R. Amode, S. Brent, W. Spooner, E. Kulesha, A. Yates, and P. Flicek, "Ensembl comparative genomics

resources.," *Database (Oxford)*, vol. 2016, p. bav096, 2016.
[27] C. Kutter, S. Watt, K. Stefflova, M. D. Wilson, A. Goncalves, C. P. Ponting, and D. T. Odom, "Rapid Turnover of Long Noncoding RNAs and the Evolution of Gene Expression," *PLoS Genet*, vol. 8, no. 7, p. e1002841, Jul. 2012.
[28] C. P. Ponting, P. L. Oliver, and W. Reik, "Evolution and functions of long noncoding RNAs.," vol. 136, no. 4, pp. 629–641, Feb. 2009.
[29] A. Smit, R. Hubley, and P. Green, "RepeatMasker Open-3.0.," 1996.
[30] U. A. Ørom and R. Shiekhattar, "Long noncoding RNAs usher in a new era in the biology of enhancers.," vol. 154, no. 6, pp. 1190–1193, Sep. 2013.

Analysis, Integration and Modeling of Cell Clustering Results in High-Dimensional Cytometry Data

Guillaume GAUTREAU, David PEJOSKI, Roger LE GRAND, Antonio COSMA, Anne-Sophie BEIGNON and Nicolas TCHITCHEK
CEA - Université Paris Sud 11 - INSERM U1184, Immunology of viral infections and autoimmune diseases, Fontenay-aux-Roses, France.

Corresponding author: nicolas.tchitchek@gmail.com

Reference paper: **Gautreau *et al.* (2016) SPADEVizR: an R package for Visualization, Analysis and Integration of SPADE results. *Bioinformatics*. http://doi.org/10.1093/bioinformatics/btw708**

Abstract *Cytometry is an experimental technique measuring cell marker expressions at the single cell level. The recent increase in the number of markers simultaneously measurable has led to the development of new automatic gating algorithms. Especially, the SPADE algorithm has been proposed as a novel way to identify clusters of cells having similar phenotypes. While SPADE or other cell clustering algorithms are powerful approaches, complementary analysis features are needed to characterize better the identified cell clusters.*

We have developed SPADEVizR, an R package designed for the visualization, analysis, and integration of cell clustering results. The available statistical methods allow highlighting cell clusters with relevant biological behaviors or integrating them with additional biological variables. Moreover, several visualization methods are available to better characterize the cell clusters. SPADEVizR can also generate mathematical models to predict biological variables, based on the cell cluster abundances.

These analysis features are essential to interpret properly the behaviors and phenotypes of the identified cell clusters.

Keywords Mass Cytometry, Automatic Gating, Statistics, Visualization, Predictive models

1 Introduction

Cytometry is an experimental technique used to characterize cell properties at the single cell level. Thanks to mass cytometry, the number of simultaneously measurable cell markers has increased up to 50 [1]. This increase of measurable cell markers has led to the development of new automatic gating algorithms to identify group of cells, also named cell clusters, having similar expressions for selected markers.

The SPADE algorithm [2] was developed to identify cell clusters in the context of mass cytometry data. SPADE is a hierarchical clustering-based algorithm combined to a density-based down-sampling procedure.

While SPADE is a powerful approach, the interpretation of the behaviors or phenotypes of the identified cell clusters can be challenging, in particular in the scope of a whole dataset. For instance, SPADE has no methods allowing to highlight cell clusters with a cell abundance statistically different between two biological conditions or associated with an additional biological variable. Moreover, SPADE lacks of visualization methods to deeply characterize the phenotypes of the cell clusters in the whole dataset.

We have developed SPADEVizR, an R package to visualize, analyze and integrate results provided by SPADE. This package extends the original SPADE outputs with techniques such as volcano plots, streamgraphs, parallel coordinates, heatmaps, or distograms. Moreover, several statistical methods allow the identification of clusters with important biological behaviors. SPADEVizR also has features allowing the quantification and the visualization of the quality of clustering results and can be used with results generated by algorithms different from SPADE.

We illustrated the capabilities of SPADEVizR in the context of a vaccine study to identify new B cell populations altered by MVA immunizations in macaques [3].

2 Statistical methods

SPADEVizR allows the identification of three types of relevant cell clusters obtained from automatic gating algorithms. Abundant Clusters (AC) correspond to clusters having a cell abundance statistically greater than

a specific threshold for a given set of samples, identified using one-sample t-tests. Differentially Abundant Clusters (DAC) correspond to clusters having a cell abundance statistically different between two biological conditions, identified using two-samples t-tests. Correlated Clusters (CC) correspond to clusters having a cell abundance statistically correlated with an additional biological variable, identified using the Pearson or Spearman coefficients of correlation. These clusters can be visualized using scatter plot or volcano plot (Fig. 1A) representations.

3 Visualization methods

Boxplot (Fig. 1B) and kinetic representations available in SPADEVizR allows efficient visualizations and comparisons of cluster abundances between different samples and conditions. Moreover, streamgraph representations can display simultaneously absolute and relative cell abundances for a set of clusters (Fig. 1C).

Phenotypical characterization of the cell clusters can be performed using categorical heatmaps or parallel coordinates (Fig. 1D). While heatmaps provide global overviews, parallel coordinates provide more details by highlighting the homogeneity of marker expressions between the samples. SPADEVizR can generate multidimensional scaling representations to visualize the similarities between samples or clusters.

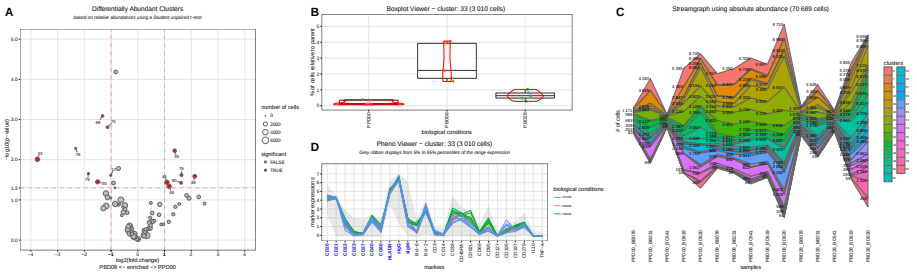


Fig. 1. Selected visualization representations available in SPADEVizR. (A) Volcano plot showing Differentially Abundant Clusters (DAC). (B) Boxplot showing the cell abundances for a given cluster in each sample and each condition. (C) Streamgraph showing absolute and relative abundances for a set of clusters across all the samples. (D) Parallel coordinates showing the phenotype of a given cluster.

4 Conclusion

SPADEVizR constitutes a powerful approach for interpreting clustering results from several automatic gating algorithms. The available methods are very valuable to analyze properly high-dimensional cytometry data.

Acknowledgements

French government program: “Investissement d’avenir: Equipements d’Excellence” (EQUIPEX) - 2010 FlowCyTech, Grant number: ANR-10-EQPX-02-01. Grant sponsor: Infrastructures Nationales en Biologie et Santé (INBS) - 2011 Infectious Disease Models and Innovative Therapies (IDMIT); Grant number: ANR-11-INBS-0008.

References

[1] Sean C Bendall, Erin F Simonds, Peng Qiu, El-ad D Amir, Peter O Krutzik, Rachel Finck, Robert V Bruggner, Rachel Melamed, Angelica Trejo, Olga I Ornatsky, Robert S Balderas, Sylvia K Plevritis, Karen Sachs, Dana Pe’er, Scott D Tanner, and Garry P Nolan. Single-cell mass cytometry of differential immune and drug responses across a human hematopoietic continuum. *Science (New York, N.Y.)*, 332(6030):687–96, May 2011.

[2] Peng Qiu, Erin F Simonds, Sean C Bendall, Kenneth D Gibbs, Robert V Bruggner, Michael D Linderman, Karen Sachs, Garry P Nolan, and Sylvia K Plevritis. Extracting a cellular hierarchy from high-dimensional cytometry data with SPADE. *Nat. Biotechnol*, 29(10):886–91, October 2011.

[3] David Pejoski, Nicolas Tchitchek, André Rodriguez Pozo, Jamila Elhmouzi-younes, Rahima Yousfi-bogniaho, Christine Rogez-kreuz, Pascal Clayette, Yves Lévy, Antonio Cosma, Roger Le Grand, and Anne Sophie Beignon. Identification of Vaccine-Altered Circulating B Cell Phenotypes Using Mass Cytometry and a Two-Step Clustering Analysis. *J. Immunol*, 196:4814–4831, 2016.

Conférence invitée

Tobias Marschall

Algorithms for Computational Genomics, Max-Planck-Institut für Informatik, Saabrücken, Germany

A Guided Tour to Computational Haplotyping

Humans and many other species are diploid. Every individual inherits two versions of each autosomal chromosome, called haplotypes, one from its mother and one from its father. Moving from (sequences of) genotypes to haplotypes is known as phasing or haplotyping. The knowledge of haplotypes is critical for addressing a variety of important questions in fundamental and clinical research. In this talk, I will highlight both algorithmic and experimental aspects of reconstructing haplotypes, with a special emphasis on recent technological advancements and their impact on the computational problems to be solved. I will briefly touch on population-based and pedigree-based phasing method, but will mostly focus on direct experimental methods that allow to reconstruct haplotypes for single individuals. Haplotype reconstruction from sequencing reads is most commonly formalized as the Minimum Error Correction (MEC) problem. Recent advances on fixed-parameter tractable (FPT) algorithm allow us to (quickly) solve practically relevant instances of this NP-hard problem optimally. I will present experimental results from five different platforms (PacBio, Oxford Nanopore, Hi-C, StrandSeq, and 10X Genomics) and highlight how combinations of these technologies allow to accurately reconstruct dense chromosome-length human haplotypes at manageable costs.

GenomeOnRails: depicting microbial species diversity via a pangenome graph

Guillaume Gautreau[1,2,3,4], Rémi Planel[1,2,3,4], Amandine Perrin[5,6], Marie Touchon[5,6], Eduardo Rocha[5,6], Christophe Ambroise[2,7,8], Catherine Matias[9], Stéphane Cruveiller[1,2,3,4], Claudine Medigue[1,2,3,4] and David Vallenet[1,2,3,4]

[1] CEA, Genoscope, 91000 Evry, France
[2] Université d'Evry, 91000 Evry, France
[3] CNRS-UMR-8030, 91000 Evry, France
[4] Université Paris-Saclay, 91000 Evry, France
[5] Microbial Evolutionary Genomics Unit, Institut Pasteur, 75724 Paris, France
[6] CNRS-UMR-3525, 75015 Paris, France
[7] Laboratoire de Mathématiques et Modélisation d'Evry (LaMME), 91000 Evry, France
[8] UMR-CNRS-8071, 91000 Evry, France
[9] Laboratoire de Probabilités et Modèles Aléatoires (LPMA), 75252 Paris, France

Corresponding Author: ggautrea@genoscope.cns.fr

Comparative genomics approaches in microbiology now use thousands of genomes to analyze a given species in different environmental or medical contexts. By collecting and comparing these genomic sequences, many studies are focused on the overall gene content of a species (*i.e.* the pangenome) 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 (*i.e.* the mobilome). As recently suggested [1], a consensus representation of multiple genomes would provide a better analysis framework than using individual reference genomes. Here, we introduce an extension of this concept, giving it a formal mathematical representation using a graph model built up from genes clustered into families.

Pangenomes are generally stored in a binary matrix denoting the presence or absence of each gene family across organisms. However, this structure does not handle the genomic organization of gene families in each organism. In our approach, we propose a graph model where nodes represent families and edges chromosomal neighborhood information. Indeed, it is known that core gene families share conserved organizations whereas variable regions are rather randomly distributed along genomes.

Based on this data structure, our method classifies gene families through an Expectation/Maximization algorithm based on Bernoulli mixture model. Moreover, in order to take in account the genomic context of gene families, we smooth the classification with neighborhood information using Markov random field model [2]. This approach splits pangenomes in three groups: (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 to environmental adaptation capabilities; (3) *cloud genome*, genes found at very low frequency.

Pangenomics is a relevant paradigm for very large scale comparative genomics. Further development of this tool should provide solid bases for efficient study of the dynamics of pangenome species and almost exhaustive references for metagenomic studies.

References

[1] Chan A. P., Sutton G., DePew J., Krishnakumar R., Choi Y., Huang X-Z., Beck E., Harkins D. M., Kim M., Lesho E. P., Nikolich M. P. and Fouts D. E. A novel method of consensus pan-chromosome assembly and large-scale comparative analysis reveal the highly flexible pan-genome of Acinetobacter baumannii. *Genome Biology*, 2015.

[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.

Assembly of heterozygous genomes with high order De Bruijn graph

Antoine LIMASSET[1], Camille MARCHET[1], Pierre PETERLONGO[1] and Jean-Francois FLOT[2]
[1] IRISA, 263 Avenue Général Leclerc, 35000 Rennes, France
[2] ULB, Avenue Franklin Roosevelt 50, 1050 Bruxelles, Belgium

Corresponding author: antoine.limasset@gmail.com

1 Introduction

Heterozygous genome assembly constitutes a complex task for which no satisfying solution exists at the present time. Phasing chromosomes in diploid or polyploid species or in metagenomes is still an open problem, despite the emergence of new long-read technologies and other dedicated approaches. In particular, intra-species or inter-species variations are usually discarded and/or result in highly fragmented assemblies. Several approaches tried to produce phased contigs using regular or paired-end short reads. On such genomes, String graph approaches like MIRA[1] produce large contigs but are intractable on large genomes. De Bruijn graph approaches (Spades[2], platanus[3], discovarDenovo[4]) are limited to very high or very low heterozygosity rates since they produce unsactisfactory fragmented contigs in the intermediary rates.

2 BWISE, a high order De Bruijn graph assembler

In this work we propose a novel de Bruijn graph-based assembler called BWISE that performs the construction of a very high order De Bruijn graph. This assembly fully takes advantage of read lengths and paired-end relationships between reads. Thereby correct paths are allowed to be determined resolving haplotypes and genomic repeats.

Results on simulated datasets show that BWISE is able to produce order of magnitude longer contigs than state of the art methods on highly heterozygous genomes. We also show that BWISE is comparable to state of the art assemblers for "regular" haploid genomes, and has the potential to scale up to very large genomes as the human one.

Beyond paired-end reads, the proposed framework allows in principle the integration of long-range information (mate pairs, 3Cseq/Hi-C, 10X, Single Molecule Real-Time (SMRT) as well as Nanopore reads) to determine accurate long paths in the assembly graph, with the ultimate goal of generating a single high-quality contigs per chromosome.

References

[1] Bastien Chevreux, Thomas Wetter, Sándor Suhai, et al. Genome sequence assembly using trace signals and additional sequence information. In *German conference on bioinformatics*, volume 99, pages 45–56. Heidelberg, 1999.

[2] Anton Bankevich, Sergey Nurk, Dmitry Antipov, Alexey A Gurevich, Mikhail Dvorkin, Alexander S Kulikov, Valery M Lesin, Sergey I Nikolenko, Son Pham, Andrey D Prjibelski, et al. Spades: a new genome assembly algorithm and its applications to single-cell sequencing. *Journal of computational biology*, 19(5):455–477, 2012.

[3] Rei Kajitani, Kouta Toshimoto, Hideki Noguchi, Atsushi Toyoda, Yoshitoshi Ogura, Miki Okuno, Mitsuru Yabana, Masayuki Harada, Eiji Nagayasu, Haruhiko Maruyama, et al. Efficient de novo assembly of highly heterozygous genomes from whole-genome shotgun short reads. *Genome research*, 24(8):1384–1395, 2014.

[4] Neil I Weisenfeld, Shuangye Yin, Ted Sharpe, Bayo Lau, Ryan Hegarty, Laurie Holmes, Brian Sogoloff, Diana Tabbaa, Louise Williams, Carsten Russ, et al. Comprehensive variation discovery in single human genomes. *Nature genetics*, 46(12):1350–1355, 2014.

Rapid spectra comparison with data-mining algorithms: accessing post translational modification profiles on a sample scale

Matthieu DAVID[1,2], Guillaume FERTIN[1], Hélène Rogniaux[2] and Dominique TESSIER[2]
[1] LS2N UMR CNRS 6004, Université de Nantes, F-44300, Nantes, France
[2] INRA UR1268 Biopolymères Interactions Assemblages, F-44316 Nantes, France

Corresponding author: matthieu.david@univ-nantes.fr

Reference paper: **David *et al.* (2016)** ***International Workshop on Algorithms in Bioinformatics*****.** **https://doi.org/10.1007/978-3-319-43681-4_6**

Abstract *Peptide identification from mass spectrometry experiments remains a challenging task, often leading to numerous missing or erroneous assignments. Because of their prohibitive computational cost, almost all peptide identification algorithms reduce their search space and in consequence, limit their ability to identify peptides displaying post-translational modifications. To resolve this difficulty, we developed a novel algorithm able to handle the computational complexity of all-to-all comparisons of spectra in the context of large volumes. We illustrated the interest of our approach with results obtained from the analysis of a public dataset and showed its capability to resolve the drawbacks usually accompanying spectra identifications without any a priori filter (e.g. computation time and decreased sensibility).*

Keywords Proteomics, Mass Spectrometry, Spectra Comparisons, Algorithms, Data Mining

1 Introduction

Tandem mass spectrometry is used to identify the proteins present in a mixture. To this end, peptides obtained from the digestion of the proteins by an enzyme are fragmented into a mass spectrometer. All the masses of the fragments are measured, and this series of masses generates an *experimental mass spectrum*.
The protein identification process relies, in a first step, on the interpretation of experimental spectra in terms of peptides. In a second step, identified peptides are used to infer the list of proteins present in the mixture. In order to interpret spectra as peptides, the most widely used methods compare experimental spectra with *theoretical spectra* inferred from in silico digestion of a protein database. A scoring function is used to evaluate the similarity between pairs of spectra, whose definition varies depending on the method; however, every such scoring function is based on the count of shared masses between spectra, as a high number of shared masses is largely accepted to lead to more reliable identifications. The number of shared masses between two given spectra is the scoring function that we used in this study.
Currently, sensibility and reliability of the protein identification process remain limited. In a classical tandem mass spectrometry experiment, up to 75% of the spectra are indeed still unassigned [1]. Achieving pairwise comparisons of tens of thousands experimental spectra against hundreds of thousands of theoretical spectra is impractical. In order to avoid an explosion of the computation time, each experimental spectrum is only compared with theoretical spectra exhibiting almost the same mass, thus reducing drastically the search space. However, since proteins display a high rate of post-translational modifications (PTMs), and because these PTMs alter the mass of the peptides, this search space reduction hinders the peptide identification process – the subset of theoretical spectra considered for identifying the peptide may not contain the correct spectrum. PTMs are therefore incriminated as the most frequent cause for missing assignments [2]. Even though traditional peptide identification engines may add some PTMs in their searches, considering too many PTMs leads once again to excessive computation time. Recent Open Modification Search (OMS) algorithms [3] resolve the loss of identifications due to modified peptide masses with a loosened mass filter. The processing time however remains a significant bottleneck, associated to a degradation of results quality with more false positives and false negatives [4].
This work proposed a novel OMS approach inspired from data-mining techniques, providing access to the PTMs profiles of the sample while avoiding the computation time explosion pitfall.

2 Methodology

We designed a new data-structure, *SpecTrees*, to identify spectra rapidly without a priori filtering. *SpecTrees* contains by construction the number of shared masses between any two spectra. Inherent property of the data layout, this information is however distributed inside the structure and requires retrieval through the specific procedure *SpecXtract*. Let s_e (resp. s_t) be an experimental (resp. theoretical) spectrum. For each (s_e, s_t) displaying a sufficiently high number of shared masses, a supplementary module analyzes this pair of spectra in order to identify peptides, possibly containing PTMs. We used a traditional statistical validation to ensure the results correctness, by means of evaluation of the False Discovery Rate (FDR).
Since the reference paper cited above was published, we extended our tests to larger datasets. We downloaded a tandem mass spectrometry dataset from the PRIDE database (PXD001428) and obtained an experimental dataset containing $37,703$ spectra. The protein database is originated from the Ensemble genome assembly and contains $500,685$ theoretical spectra after replication of the digestion enzyme behavior.

3 Results

Altogether, *SpecTrees* and *SpecXtract* compute the number of shared masses between any pair of experimental and theoretical spectra (s_e, s_t). We achieved on the larger datasets a running time under 10 minutes for different sets of parameters, clearly outperforming traditional identification tools or other OMS approaches. In-memory occupation of the program did not exceed 3 GigaBytes, and compatibility with the current throughput of proteomic analyzes is therefore ensured.
Our experiment demonstrated the ability of our software to identify $11,404$ spectra, approximately 28% of the total, with a false detection rate of about 1%. Among those identifications, some classical cases of missed-cleavage (absence of digestion from the enzyme at a given location between two peptides), carbamylation, deamidation, dioxidation, formylation, oxidation and isotopic peptides were found, therefore validating the consistency of our approach. Rare modifications were also highlighted with a sufficient confidence, some of which not yet described in PTMs databanks. Our approach additionally confers more reliability to the FDR for modified peptides using search space stratification, where traditional algorithms fail to provide an homogeneous FDR [5].

4 Conclusion

The *SpecTrees* data-structure is designed to compare important collections of spectra. Without any preliminary mass filter in spectra comparisons, *SpecTrees* can highlight proteins with original PTMs useful to direct further biological experiments. Moreover, flexible design enables the use of our method to detect peak patterns within any set of masses. Finally, the execution time of the method is compatible with a routine use in mass spectrometry laboratories.

Acknowledgements

This project is partly funded by the Région Pays de la Loire (France) GRIOTE program (2013-2018).

References

[1] J. Griss, Y. Perez-Riverol, S. Lewis, D. L. Tabb, J. A. Dianes, N. Del-Toro, M. Rurik, M. W. Walzer, O. Kohlbacher, H. Hermjakob, R. Wang, and J. A. Vizcaino. Recognizing millions of consistently unidentified spectra across hundreds of shotgun proteomics datasets. *Nat. Methods*, 13(8):651–656, Aug 2016.

[2] B. Bogdanow, H. Zauber, and M. Selbach. Systematic Errors in Peptide and Protein Identification and Quantification by Modified Peptides. *Mol. Cell Proteomics*, 15(8):2791–2801, Aug 2016.

[3] E. Ahrne, M. Muller, and F. Lisacek. Unrestricted identification of modified proteins using MS/MS. *Proteomics*, 10(4):671–686, Feb 2010.

[4] S. Na, N. Bandeira, and E. Paek. Fast multi-blind modification search through tandem mass spectrometry. *Mol. Cell Proteomics*, 11(4):M111.010199, Apr 2012.

[5] G. Hart-Smith, D. Yagoub, A. P. Tay, R. Pickford, and M. R. Wilkins. Large Scale Mass Spectrometry-based Identifications of Enzyme-mediated Protein Methylation Are Subject to High False Discovery Rates. *Mol. Cell Proteomics*, 15(3):989–1006, Mar 2016.

Conférence invitée

Julio Saez-Rodriguez

RWTH, Aachen University, Germany and EMBL-EBI

Network Models to Understand and Combat Cancer: from Clinical Genomics to Biochemical Modelling

Large-scale genomic studies are providing unprecedented insights into the molecular basis of cancer, but it remains challenging to leverage this information for the development and application of therapies. We have performed an integrated analysis of the molecular profiles of 11,215 primary tumours and 1,001 cancer cell lines, along with the response of the cell lines to 265 anti-cancer compounds. This analysis finds alterations in tumours that can confer drug sensitivity or resistance, and sheds light on which data types are most informative to prioritize treatment. Integration of this data with various sources of prior knowledge, in particular signaling pathways and transcription factors, points at molecular processes involved in resistance mechanisms, and offer hypotheses for novel combination therapies. Our own analysis as well as the results of a crowdsourcing effort (DREAM challenge) reveals that prediction of drug efficacy is far from accurate, implying important limitations for personalised medicine. I will argue than an important missing aspect is the dynamics of signaling networks, and show how applying logic models, trained with phosphoproteomic measurements upon perturbations, can further improve our understanding of the molecular basis of drug resistance, thereby providing new treatment opportunities not noticeable by static molecular characterization.

Rare and common epilepsies converge on a shared gene regulatory network providing opportunities for novel antiepileptic drug discovery

Andrée DELAHAYE-DURIEZ[1,2,3,4], Prashant SRIVASTAVA[1,*], Kirill SHKURA[1,*], Sarah R. LANGLEY[1,5,*], Liisi LAANISTE[1], Aida MORENO-MORAL[2,5], Bénédicte DANIS[6], Manuela MAZZUFERI[6], Patrik FOERCH[6], Elena V. GAZINA[7], Kay RICHARDS[7], Steven PETROU[7,8,9], Rafal M. KAMINSKI[6], Enrico PETRETTO[2,5] and Michael R. JOHNSON[1]

[1] Division of Brain Sciences, Imperial College Faculty of Medicine, London, UK
[2] MRC Clinical Sciences Centre, Imperial College London, London, UK
[3] Université Paris 13, Sorbonne Paris Cité, UFR de Santé, Médecine et Biologie Humaine, France
[4] PROTECT, INSERM, Unversité Paris Diderot, Sorbonne Paris Cité, Paris, France
[5] Duke-NUS Medical School, 8 College Road 169857 Singapore, Republic of Singapore
[6] Neuroscience TA, UCB Pharma, S.A. Allée de la Recherche, 60 1070 Brussels, Belgium
[7] The Florey Institute of Neuroscience and Mental Health, The University of Melbourne, Parkville, Victoria 3052, Australia
[8] The Centre for Neural Engineering, The Department of Electrical Engineering, The University of Melbourne, Parkville, Victoria 3052, Australia
[9] The Australian Research Council Centre of Excellence for Integrative Brain Function, Parkville, Victoria 3052, Australia

[*] These authors participated equally to this work as joint second authors

Corresponding Authors: andree.delahaye@inserm.fr
enrico.petretto@duke-nus.edu.sg
m.johnson@imperial.ac.uk

Paper Reference: **Delahaye-Duriez *et al.* (2016) Rare and common epilepsies converge on a shared gene regulatory network providing opportunities for novel antiepileptic drug discovery. Genome Biology.**
http://dx.doi.org/ 10.1186/s13059-016-1097-7

Abstract

Background

The relationship between monogenic and polygenic forms of epilepsy is poorly understood and the extent to which the genetic and acquired epilepsies share common pathways is unclear. Here, we use an integrated systems-level analysis of brain gene expression data to identify molecular networks disrupted in epilepsy.

Results

We identified a co-expression network of 320 genes (M30), which is significantly enriched for non-synonymous *de novo* mutations ascertained from patients with monogenic epilepsy and for common variants associated with polygenic epilepsy. The genes in the M30 network are expressed widely in the human brain under tight developmental control and encode physically interacting proteins involved in synaptic processes. The most highly connected proteins within the M30 network were preferentially disrupted by deleterious *de novo* mutations for monogenic epilepsy, in line with the centrality-lethality hypothesis. Analysis of M30 expression revealed consistent downregulation in the epileptic brain in heterogeneous forms of epilepsy including human temporal lobe epilepsy, a mouse model of acquired temporal lobe epilepsy, and a mouse model of monogenic Dravet (SCN1A) disease. These results suggest functional disruption of M30 via gene mutation or altered expression as a convergent mechanism regulating susceptibility to epilepsy broadly. Using the large collection of drug-induced gene expression data from Connectivity Map, several drugs were predicted to preferentially restore the

downregulation of M30 in epilepsy toward health, most notably valproic acid, whose effect on M30 expression was replicated in neurons.

The code of R functions newly created for this study is provided as a convenient R-package at https://github.com/adelahay/BrainCell (DOI:10.5281/zenodo.164147).

Conclusions

Taken together, our results suggest targeting the expression of M30 as a potential new therapeutic strategy in epilepsy.

Keywords Epilepsy, Co-expression, Regulatory network, Protein-protein interactions, Valproic acid.

Prediction of Disease-associated Genes by advanced Random Walk with Restart on Multiplex and Heterogeneous Biological Networks

Alberto VALDEOLIVAS[1,2], Elisabeth REMY[1], Laurent TICHIT[1], Gaëlle ODELIN[2] Claire NAVARRO[2], Sophie PERRIN[2], Pierre CAU[3], Nicolas LEVY[3] and Anaïs BAUDOT[1]

[1] Aix-Marseille Université, CNRS, Centrale Marseille, I2M UMR 7373, Marseille, France.
[2] ProGeLife, 8 Rue Sainte Barbe 13001, Marseille, France.
[3] Aix-Marseille Université, INSERM, UMR_S910, Faculté de Médecine, France.

Corresponding author: alberto.valdeolivas@etu.univ-amu.fr, anais.baudot@univ-amu.fr

Abstract *Rare monogenic diseases globally affect millions of persons, but many causative genes remain to be discovered. Several computational approaches have been developed to predict disease-associated genes. Guilt-by-association strategies on protein interaction networks, in particular, postulate that proteins lying in a close network vicinity are functionally-related and implicated in similar phenotypes.*

However, current network approaches are limited as they do not exploit the richness of biological networks, which are both multiplex (i.e., containing different layers of physical and functional interactions between genes and proteins), and heterogeneous (i.e., containing both interactions between genes/proteins, and interactions between diseases). In the present study, we extended the Random Walk with Restart algorithm to leverage these complex biological networks.

We compared our algorithm to classical random walks thanks to a leave-one-out strategy. The Random Walk with Restart on multiplex and heterogeneous networks takes advantage of data pluralism and shows increased performances to predict known disease-associated genes. We finally applied it to predict candidate genes for the Wiedemann-Rautenstrauch Syndrome.

Keywords Random Walks with Restart, Biological Networks, Multiplex and Heterogeneous Networks, Disease-Gene Prioritization.

1 Introduction

Rare monogenic diseases are often opposed to common diseases, but they jointly affect millions of persons. Overall, the causative gene(s) are often unknown, many patients remain undiagnosed, and no treatment exists for most of them. The disease phenotypes are resulting not from perturbations of isolated genes or proteins, but of complex networks of molecular interactions [1,2,3]. Proteins, for instance, do not act in isolation, but rather interact with each other to perform their functions in signalling pathways or metabolic reactions. Thanks to the scaling of the experimental techniques allowing interaction discovery, recent years have witnessed the accumulation of interaction datasets. For instance, protein-protein interactions (PPI) are nowadays screened at the proteome scale revealing thousands of physical interactions between proteins. Interaction data are commonly represented as networks, in which the nodes correspond to the genes or proteins, and the edges to their interactions.

In this context, network-mining approaches are applied to study human diseases, and in particular rare monogenic diseases. The rationale underlying these network approaches for human diseases is the clustering of proteins participating to the same cellular functions or biological processes in close network vicinity. Consequently, mutated genes coding for network-related proteins will lead to the same or similar phenotypes [4]. Following this idea, the mapping of the human protein-protein interactome network, about 10 years ago, was used to reveal new disease-associated genes, or allowed prioritizing candidate disease genes [5,6,7]. It also unveiled unsuspected interactions between disease-causing proteins, such as between proteins coded by genes mutated in ataxias [8]. More globally, interaction networks can also help deciphering the etiology and physiopathology of diseases [3], and their comorbidity relationships through their distances in molecular networks [9].

Among the various network-based study of human genetic diseases, Random Walk with Restart (RWR) appears as one of the state-of-the-art guilt-by-association approaches to predict new candidate disease genes. It was initially applied to explore the surroundings of disease-associated protein *seed(s)* in a PPI network. Every

protein in the global network is ranked according to its affinity to the seed(s), thereby allowing the prioritization of new candidate disease proteins [10]. The RWR algorithm was then extended, in particular to leverage phenotypic information [11,12,13,14]. For instance, Li and Patra RWR algorithm [11] considers jointly a PPI network and a network of phenotypic similarities between diseases. The two networks are connected by bipartite protein-disease associations, and form a heterogeneous network, i.e. a network containing nodes and edges of different nature connected by bipartite associations.

However, a common feature and limitation of these approaches is that they ignore the rich variety of information on physical and functional relationships between genes and proteins. Indeed, not only PPI are nowadays described on a large-scale: affinity purification followed by mass-spectrometry experiments inform on the *in vivo* molecular complexes, pathways interaction data are cured and stored in dedicated databases. In addition, functional interactions can be derived, for instance by constructing a co-expression network from transcriptomics expression data. Overall, the exploitation of this diversity of interaction data is lagging behind.

Sets of networks sharing the same nodes, but in which edges belong to different categories or represent interactions of different nature, are known as multiplex (aka multi-layer or multi-slice) networks [15]. In a biological multiplex network, each layer contains a different category of physical and functional interactions between genes or proteins. The combination of the different interaction sources, each having its own features and bias, provides a complementary view on genes and protein cellular functioning.

We present here the extension of the RWR algorithm to multiplex and heterogeneous biological networks (RWR-MH). We demonstrate the increased performance of this algorithm when compared with classical and current RWR approaches. Finally, as a real-case biological example, we applied the RWR-MH algorithm to predict candidate genes for the Wiedemann-Rautenstrauch syndrome (WRS), whose responsible gene(s) remain unknown.

2 Methods

We constructed a multiplex network as described in [16], but updated from downloads on November 2016 (Tab 1). The multiplex network is composed of 3 layers of physical and functional interactions between genes and proteins: protein-protein interactions (PPI), Pathway interactions extracted from pathway databases and Co-Expression interactions derived from RNA-seq expression data. The network nodes correspond equally to genes or proteins.

Additionally, we built a disease-disease similarity network, in which the edges between 2 diseases correspond to significant phenotype similarities. Briefly, we retrieved diseases and their associated phenotypes from the Human Phenotype Ontology (HPO) [17]. We then computed the phenotype similarity between diseases, by measuring the relative information content of the common phenotypes of every disease pair. We thereby assumed that rare phenotypes in the HPO database are more informative than frequent ones, as proposed by [18]. Finally, the disease-disease similarity network is constructed by linking every disease to its 5 most similar diseases according to their phenotype similarity scores, as proposed in [11].

The multiplex and heterogeneous network is constructed by linking every layer of the multiplex with the disease-disease similarity network using known gene-disease bipartite associations extracted from OMIM [19].

Network	**Number of nodes**	**Number of edges**
PPI	12 621	66 971
Pathways	10 534	254 766
Co-expression	10 458	1 337 347
Disease-disease similarity	6 947	28 246

Tab. 1. Size of the networks used in this work.

We extended the RWR algorithm to consider multiplex and heterogeneous networks (RWR-MH). In a nutshell, starting from initial seed node(s), the RWR progresses following the graph topology, with a non-zero probability to jump back to the initial seed node(s) at each step. After a sufficiently large number of iterations, RWR reaches a stationary state. In this stationary state, each nodes is associated to a score reflecting its proximity or pertinence with respect to the initial seed(s). Our extended RWR-MH algorithm has the capacity to explore one network layer, but also to jump between the different layers of the multiplex, because the same

nodes are present in the different layers. In addition, it can jump to the disease-disease similarity network thanks to the gene-disease bipartite associations. At every time, the walk can return to the initial protein and/or disease seed node(s), with a defined probability that we set to 0.7, as in previous studies [10,11,12].

We applied a leave-one-out cross-validation (LOOCV) strategy to compare RWR-MH with other RWR approaches. For every disease in OMIM [19] associated to 2 genes or more, each disease-associated gene is removed one-by-one (we will later refer to this removed gene as the left-out gene). The remaining disease-associated genes and the disease itself are used as seed nodes, and the RWR algorithms are applied. The Cumulative Distribution Functions (CDFs) are used to evaluate and compare the performances of the different approaches. They display the percentage of left-out genes that are ranked within the top k genes.

3 Results

3.1 RWR-MH outperforms current RWR approaches

We first compared different algorithms: i) the classical RWR on a monoplex PPI network, ii) the RWR on a heterogeneous network (built with the PPI and the disease-disease similarity network, RWR-H, as proposed by [11]), iii) the RWR on a multiplex network (composed of 3 layers, namely PPI, pathways and co-expression, RWR-M), and finally iv) the RWR on the multiplex-heterogeneous network (RWR-MH).

We applied a LOOCV strategy to evaluate the performances of the different RWR algorithms (Methods) for the prediction of known disease-associated genes (Fig 1). The multiplex RWR-M shows a superior performance than the classical RWR on a monoplex network. It is however comparable to the heterogeneous RWR-H. The multiplex and heterogeneous RWR-MH is remarkably better, since more than 45% of the left-out genes are ranked within the top 20.

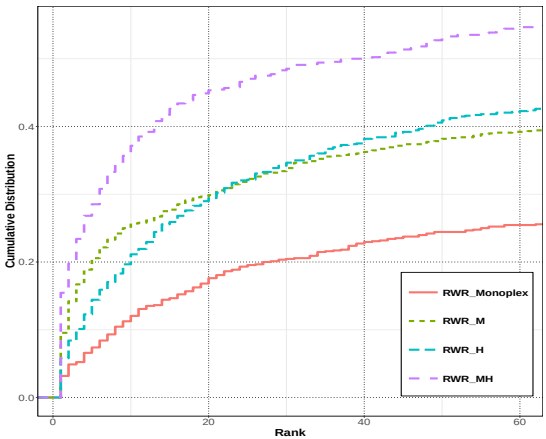


Fig. 1. Cumulative Distribution Functions of the rank of each left-out gene retrieved by the LOOCV strategy. The algorithms are the classical RWR applied to the PPI monoplex network, the RWR-M applied to the 3-layer multiplex network, the RWR-H applied to the heterogeneous network built from the PPI and the disease-disease similarity networks, and the RWR-MH applied to the multiplex-heterogeneous network.

3.2 RWR-MH prediction of candidate genes for the Wiedemann-Rautenstrauch Syndrome

The Wiedemann-Rautenstrauch Syndrome (WRS; MIM code: 264090), also called neonatal progeroid syndrome, is a disorder characterized by intrauterine growth retardation with subsequent failure to thrive and short stature [20]. In addition, patients display a progeroid appearance at birth and during the infancy, decreased subcutaneous fat, hypotrichosis and macrocephaly [21]. Only a few case reports have been documented, and no gene(s) has been described as causative of the syndrome yet.

To illustrate our approach, we applied our RWR-MH algorithm using as seed the WRS disease node. For visualization purposes, we displayed only the top 25 ranked diseases and top 25 ranked genes scored by the RWR-MH algorithm (Fig 2).

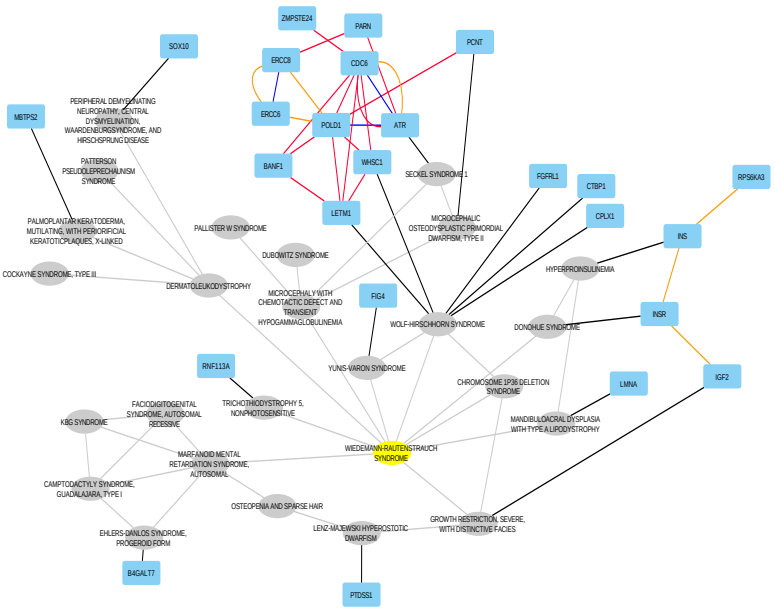


Fig. 2. Multiplex and Heterogeneous Network linking the top 25 genes and top 25 diseases obtained when the RWR-MH algorithm is applied using WRS as seed (yellow node). Grey elliptical nodes represent diseases, while turquoise rectangles represent genes. Black edges account for the bipartite gene-disease associations; Grey edges are the similarity links in the disease-disease similarity network; Blue edges correspond to PPI interactions; Red edges represent co-expression relationships; Orange edges represent pathway interactions.

We predicted the top scored genes as top candidates for being involved in WRS. Many of them, such as *FIG4*, *RNF113A* or *LMNA*, are implicated in diseases directly connected to WRS from phenotypic similarities. For instance, mutations in *LMNA* are responsible for Hutchinson-Gilford progeria syndrome (HGPS; MIM code: 176670) and other premature aging syndromes such as Mandibuloacral Dysplasia with type A Lipodystrophy (MAD-A; MIM code: 248370). However, no mutations were found targeting *LMNA* sequence by gene sequencing analyses of WRS patients [21,22]. Additionally, the RWR-MH algorithm evidenced *ZMPSTE24*, which is known to be responsible of severe progeroid syndromes such as restrictive dermopathy (RD, MIM code: 275210) [23]. This peptidase acts during the post-translation modifications of the prelamin A, coded by *LMNA*, to undergo the complete maturation to lamin A. The direct interaction between the products of *LMNA* and *ZMPSTE24* is missing in the databases we used to construct our multiplex network. The *ZMPSTE24* nodes is however retrieved through different trajectories in the random walk. Once again, no mutations were found in the *ZMPSTE24* gene among the 5 WRS patients [22].

An interesting result is the small subnetwork composed of the genes *IGF2*, *INS*, *INSR* and *RPS6KA3*, which all participate to the insulin pathway. We retrieved these genes as top candidates due to their associations with two different diseases linked to WRS. This pathway is suspected to play a role in WRS [22]. Similarly, a cluster of proteins related to the cell cycle and DNA repair is connected to WRS through the Wolf-Hirschhorn syndrome (MIM code: 194190). DNA repair defects are also suspected to be involved in WRS [22]. The next step will be to validate these predictions, for instance using exome-sequencing data. Overall, our extended guilt-by-association RWR-MH algorithm could be integrated in analysis pipelines to help predicting candidate genes for rare diseases.

4 Discussion

Random Walk with Restart (RWR) is one of the state-of-the-art guilt-by-association approaches to prioritize candidate disease genes. We here extended the classical RWR algorithm in order to navigate multiplex and heterogeneous networks. We also demonstrated the increased performance of the RWR multiplex and heterogeneous strategy by leave-one-out cross validations. This improvement is due to the ability of our algorithm to extract and integrate the information from many interaction sources. Other types of networks could be integrated in the future, for instance to include interactions with non-coding RNAs. In addition, it would be interesting to explore the impact on the results of different disease-disease network topologies, i.e. taking a different criteria to build the disease-disease phenotype similarity network.

RWR has been mainly employed in biology to predict disease-associated genes [10,11,12,14,13]. But it has also been applied to address other biological problems, such as the prediction of drug-target potential interactions [24], the identification of clusters from PPI networks [25], or the prediction of adverse drug reactions [26]. Multiplex approaches are likely to boost the results of all these different applications, and could also be adapted to study cellular functioning as a whole.

The next step will be to resolve the degree bias of the algorithm. Indeed, RWR algorithms and other network propagation methods are biased towards networks hubs [27]. Therefore, poorly-connected and not well-characterized proteins, which can also be related to diseases, are hard to detect. Biased random walks can tackle this issue by relating the probability of transition to the degree of the nodes [28].

Acknowledgements

This work was supported by JOBIM 2017.

References

[1] Hannah Carter, Matan Hofree, and Trey Ideker. Genotype to phenotype via network analysis. *Current opinion in genetics & development*, 23(6):611–21, dec 2013.

[2] Marc Vidal, Michael E Cusick, and Albert-László Barabási. Interactome networks and human disease. *Cell*, 144(6):986–98, mar 2011.

[3] Trey Ideker and Roded Sharan. Protein networks in disease. *Genome research*, 18(4):644–52, apr 2008.

[4] M Oti and H G Brunner. The modular nature of genetic diseases. *Clinical Genetics*, 71(1):1–11, jan 2007.

[5] Miguel Angel Pujana, Jing-Dong J Han, Lea M Starita, Kristen N Stevens, Muneesh Tewari, Jin Sook Ahn, Gad Rennert, Víctor Moreno, Tomas Kirchhoff, Bert Gold, Volker Assmann, Wael M Elshamy, Jean-François Rual, Douglas Levine, Laura S Rozek, Rebecca S Gelman, Kristin C Gunsalus, Roger a Greenberg, Bijan Sobhian, Nicolas Bertin, Kavitha Venkatesan, Nono Ayivi-Guedehoussou, Xavier Solé, Pilar Hernández, Conxi Lázaro, Katherine L Nathanson, Barbara L Weber, Michael E Cusick, David E Hill, Kenneth Offit, David M Livingston, Stephen B Gruber, Jeffrey D Parvin, and Marc Vidal. Network modeling links breast cancer susceptibility and centrosome dysfunction. *Nature genetics*, 39(11):1338–49, nov 2007.

[6] Insuk Lee, U Martin Blom, Peggy I Wang, Jung Eun Shim, and Edward M Marcotte. Prioritizing candidate disease genes by network-based boosting of genome-wide association data. *Genome research*, 21(7):1109–21, jul 2011.

[7] Kasper Lage, E Olof Karlberg, Zenia M Størling, Páll I Olason, Anders G Pedersen, Olga Rigina, Anders M Hinsby, Zeynep Tümer, Flemming Pociot, Niels Tommerup, Yves Moreau, and Søren Brunak. A human phenome-interactome network of protein complexes implicated in genetic disorders. *Nature Biotechnology*, 25(3):309–316, mar 2007.

[8] Janghoo Lim, Tong Hao, Chad Shaw, Akash J Patel, Gábor Szabó, Jean-François Rual, C Joseph Fisk, Ning Li, Alex Smolyar, David E Hill, Albert-László Barabási, Marc Vidal, and Huda Y Zoghbi. A protein-protein interaction network for human inherited ataxias and disorders of Purkinje cell degeneration. *Cell*, 125(4):801–14, may 2006.

[9] Jörg Menche, Amitabh Sharma, Maksim Kitsak, Susan Dina Ghiassian, Marc Vidal, Joseph Loscalzo, and Albert-László Barabási. Uncovering disease-disease relationships through the incomplete interactome. *Science*, 347(6224), feb 2015.

[10] Sebastian Köhler, Sebastian Bauer, Denise Horn, and Peter N Robinson. Walking the Interactome for Prioritization of Candidate Disease Genes. *AJHG*, 82(April):949–958, 2008.

[11] Yongjin Li and Jagdish C. Patra. Genome-wide inferring gene-phenotype relationship by walking on the heterogeneous network. *Bioinformatics*, 26(9):1219–1224, 2010.

[12] Yongjin Li and Jinyan Li. Disease gene identification by random walk on multigraphs merging heterogeneous genomic and phenotype data. *BMC genomics*, 13 Suppl 7(Suppl 7):S27, 2012.

[13] MaoQiang Xie, YingJie Xu, YaoGong Zhang, TaeHyun Hwang, and Rui Kuang. Network-based phenome-genome association prediction by bi-random walk. *PLoS ONE*, 10(5):1–18, 2015.

[14] Zhi Qin Zhao, Guo Sheng Han, Zu Guo Yu, and Jinyan Li. Laplacian normalization and random walk on heterogeneous networks for disease-gene prioritization. *Computational Biology and Chemistry*, 57:21–28, 2015.

[15] Federico Battiston, Vincenzo Nicosia, and Vito Latora. Structural measures for multiplex networks. *Physical Review E - Statistical, Nonlinear, and Soft Matter Physics*, 89(3):1–16, 2014.

[16] Gilles Didier, Christine Brun, and Anaïs Baudot. Identifying Communities from Multiplex Biological Networks. *PeerJ*, pages 1–9, 2015.

[17] Sebastian Köhler, Sandra C. Doelken, Christopher J. Mungall, Sebastian Bauer, Helen V. Firth, Isabelle Bailleul-Forestier, Graeme C M Black, Danielle L. Brown, Michael Brudno, Jennifer Campbell, David R. Fitzpatrick, Janan T. Eppig, Andrew P. Jackson, Kathleen Freson, Marta Girdea, Ingo Helbig, Jane A. Hurst, Johanna Jähn, Laird G. Jackson, Anne M. Kelly, David H. Ledbetter, Sahar Mansour, Christa L. Martin, Celia Moss, Andrew Mumford, Willem H. Ouwehand, Soo Mi Park, Erin Rooney Riggs, Richard H. Scott, Sanjay Sisodiya, Steven Van Vooren, Ronald J. Wapner, Andrew O M Wilkie, Caroline F. Wright, Anneke T. Vulto-Van Silfhout, Nicole De Leeuw, Bert B A De Vries, Nicole L. Washingthon, Cynthia L. Smith, Monte Westerfield, Paul Schofield, Barbara J. Ruef, Georgios V. Gkoutos, Melissa Haendel, Damian Smedley, Suzanna E. Lewis, and Peter N. Robinson. The Human Phenotype Ontology project: Linking molecular biology and disease through phenotype data. *Nucleic Acids Research*, 42(D1):966–974, 2014.

[18] Sarah K Westbury, Ernest Turro, Daniel Greene, Claire Lentaigne, Anne M Kelly, Tadbir K Bariana, Ilenia Simeoni, Xavier Pillois, Antony Attwood, Steve Austin, Sjoert Bg Jansen, Tamam Bakchoul, Abi Crisp-Hihn, Wendy N Erber, Rémi Favier, Nicola Foad, Michael Gattens, Jennifer D Jolley, Ri Liesner, Stuart Meacham, Carolyn M Millar, Alan T Nurden, Kathelijne Peerlinck, David J Perry, Pawan Poudel, Sol Schulman, Harald Schulze, Jonathan C Stephens, Bruce Furie, Chris Van Geet, Augusto Rendon, Keith Gomez, Michael A Laffan, Michele P Lambert, Paquita Nurden, Willem H Ouwehand, Sylvia Richardson, and Andrew D Mumford. Human phenotype ontology annotation and cluster analysis to unravel genetic defects in 707 cases with unexplained bleeding and platelet disorders. *Genome Medicine*, 7:36, 2015.

[19] Ada Hamosh, Alan F. Scott, Joanna S. Amberger, Carol A. Bocchini, and Victor A. McKusick. Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders. *Nucleic Acids Research*, 33(DATABASE ISS.):514–517, 2005.

[20] H. V. Toriello. Syndrome of the month: Wiedemann-Rautenstrauch syndrome. *J. Med. Genet.*, pages 256–257, 1990.

[21] Aslihan Kiraz, Samim Ozen, Filiz Tubas, Yusuf Usta, Ozgur Aldemir, and Yasemin Alanay. Wiedemann-Rautenstrauch syndrome: Report of a variant case. *American Journal of Medical Genetics, Part A*, 158 A(6):1434–1436, 2012.

[22] Jia Woei Hou. Natural Course of Neonatal Progeroid Syndrome. *Pediatrics and Neonatology*, 50(3):102–109, 2009.

[23] Claire L. Navarro, Pierre Cau, and Nicolas Lévy. Molecular bases of progeroid syndromes. *Human Molecular Genetics*, 15(SUPPL. 2):151–161, 2006.

[24] Hui Liu, Mengmeng Guo, Ting Xue, Jihong Guan, and Libo Luo. Screening lifespan-extending drugs in Caenorhabditis elegans via label propagation on drug-protein networks. *BMC Systems Biology*, 10(Suppl 4), 2016.

[25] Kathy Macropol, Tolga Can, and Ambuj Singh. RRW: repeated random walks on genome-scale protein networks for local cluster discovery. *BMC Bioinformatics*, 10(1):283, 2009.

[26] Xiaowen Chen, Hongbo Shi, Feng Yang, Lei Yang, Yingli Lv, Shuyuan Wang, Enyu Dai, Dianjun Sun, Wei Jiang, K. M. Giacomini, M. Roy, R. Dumaine, A. M. Brown, E. Lounkine, L. Yang, J. Chen, L. He, L. Yang, J. B. Pan, M. Kuhn, M. Rarey, B. Kramer, T. Lengauer, G. Klebe, M. Campillos, M. Kuhn, A. C. Gavin, L. J. Jensen, P. Bork, L. Brouwers, M. Iskar, G. Zeller, V. van Noort, P. Bork, F. Napolitano, E. Bresso, M. Kuhn, M. Campillos, I. Letunic, L. J. Jensen, P. Bork, Z. L. Ji, J. X. Zhang, Z. Gao, X. Chen, Z. L. Ji, Y. Z. Chen, D. Szklarczyk, Y. Li, J. C. Patra, X. Chen, M. X. Liu, G. Y. Yan, S. Kohler, S. Bauer, D. Horn, P. N. Robinson, Q. Jiang, M. Duran-Frigola, P. Aloy, W. Jiang, M. Zhou, Y. Lv, J. Pinero, C. J. Zheng, G. Bindea, J. Turkson, R. Jove, F. A. Zouein, M. N. Richard, J. F. Deniset, A. L. Kneesh, D. Blackwood, G. N. Pierce, H. Kobori, M. Nangaku, L. G. Navar, A. Nishiyama, A. E. Cain, R. A. Khalil, and R. A. Khalil. Large-scale identification of adverse drug reaction-related proteins through a random walk model. *Scientific Reports*, 6(August):36325, 2016.

[27] Sinan Erten, Gurkan Bebek, Rob M Ewing, and Mehmet Koyutürk. DADA: Degree-Aware Algorithms for Network-Based Disease Gene Prioritization. *BioData mining*, 4(1):19, 2011.

[28] Federico Battiston, Vincenzo Nicosia, and Vito Latora. Efficient exploration of multiplex networks. *New Journal of Physics*, 18(4):043035, 2016.

CRCmapper: models of core transcriptional regulatory circuitries

Violaine Saint-André[1,2], Alexander J Federation[3], Charles Y Lin[3], Brian J Abraham[1], Jessica Reddy[1,4] , Tong Ihn Lee[1], James E Bradner[2,5]and Richard A Young[1,4]

[1] Whitehead Institute for Biomedical Research, Cambridge, Massachusetts 02142, USA
[2] Institut Curie, CNRS UMR3244/UPMC, 26 rue d'Ulm,75005 Paris, France
[3] Department of Medical Oncology, Dana-Farber Cancer Institute, Boston, Massachusetts 02215, USA
[4] Department of Biology, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139, USA
[5] Department of Medicine, Harvard Medical School, Boston, Massachusetts 02115, USA

Corresponding Author: violaine.saint-andre@curie.fr

There is considerable evidence that the control of gene expression programs is dominated by a small number of transcription factors (TFs). In embryonic stem cells and a few other cell types, these core TFs collectively regulate their own gene expression, forming an interconnected auto-regulatory loop that is considered the core transcriptional regulatory circuitry (CRC). There is limited knowledge of core TFs for most cell types. We recently discovered that genes encoding known core TFs forming CRCs are driven by super-enhancers, which provides an opportunity to systematically predict CRCs in poorly studied cell types through super-enhancer mapping. We have developed a tool, CRCmapper [1], which enables users to generate transcriptional regulatory circuitry maps for any sample for which they can provide suitable ChIP-seq data. Applying CRCmapper to a large set of human samples we have generated core circuitry models models for 75 human cell and tissue types. These core circuitry models recapitulate and expand on what on previous CRCs and should prove valuable for further investigating cell-type−specific transcriptional regulation in healthy and diseased cells.

References

[1] Saint-André V, Federation AJ, Lin CY, Abraham BJ, Reddy J, Lee TI, Bradner JE, Young RA. Models of human core transcriptional regulatory circuitries. *Genome Res*. 2016 Mar;26(3):385-96.

Network Inference of Dynamic Models by the Combination of Spanning Arborescences

Anthony COUTANT and Céline ROUVEIROL
Laboratoire d'Informatique de Paris Nord (LIPN), UMR CNRS 7030, Université Paris 13
99 avenue Jean-Baptiste Clément, 93430 Villetaneuse, France

Corresponding author: firstname.lastname@lipn.univ-paris13.fr

Abstract *In this paper, we tackle the problem of generative learning of dynamic models from "fat" time series data (high #variables/#individuals ratio), leading to a high sensitivity of learned models to the dataset noise. To overcome this problem, we propose a method computing a mixture of many highly biased but optimal spanning arborescences obtained from many perturbed versions of the original dataset, introducing variance to counterbalance the strong arborescence bias. The method is theoretically at the boundary between structure oriented Bayesian model averaging and recent work on density estimation using mixtures of poly-trees through a perturb and combine framework, transposed to a dynamic setting. In practice, preliminary results on the recent DREAM D8C1 challenge are promising.*

Keywords Network inference, Ensemble Learning, Model Averaging, Spanning Arborescences

1 Introduction

Accurate network inference, or generative structure learning, is a major problem in bioinformatics for the comprehension of systems of different kinds, such as regulatory networks [1] or cell signaling pathways [2]. Although a difficult problem in any dataset, generative structure learning on biological contexts is additionally challenged by the scarcity of the available datasets, obtaining such data often being very expensive and time consuming.

A major issue occuring in structure learning algorithms from data is the one of *data fragmentation*, where the computation of frequencies from data is not reliable enough to robustly estimate statistics and avoid overfitting. This problem occurs more particularly when the dataset is scarce and suffers from a very high variables/samples ratio, such as in many biological contexts where the number of available sets of measurements does not exceed several tens in a good situation, for a living organism made of hundreds or thousands of genes and involving many more molecules.

Biological systems are also often characterized by their dynamic properties, and many biological phenomenon under study are actually time evolving processes, e.g. the diauxic shift of the *Saccharomyces cerevisiae* yeast [3], for which available datasets can take the form of time series. Together, the data scarcity and the objective of finding dynamic models reinforce the data fragmentation issue, since considering several time points in space for each variable in the model both increases the number of actually considered variables in the network to learn, and reduces the samples size of the original dataset, due to the underlying used sliding window.

A possible strategy to prevent overfitting in such situation is to reduce the learning algorithm variance by reducing the number of possible models. This can be achieved for example by constraining the search space so that the resulting subspace have good properties, or by constraining the search algorithm so that a subset of possible models is reachable. Used alone, this strategy can find a good local, even global, optimum, but relatively to a potentally inadequate space where a good solution for the overall learning problem is unexistent.

Another possibility is to find an asymptotic structure which is the result of a consensus between different models learned from the dataset [4,5]. This way, the lack of sufficient statistics on data is partly offset by an attempt to compute sufficient statistics on potentially very noisy intermediate models.

In this article, we use both solutions. The general behaviour of the algorithm we propose is to compute an expected *composite* model using Bayesian Model Averaging [4] theory and a set of expected features, the expected edge existence, computed from the learning of many *component* models. These components are themselves highly biased models, more precisely spanning arborescences [6], with the interesting property for

these simple structure spaces, that a global optimum is computable in polynomial time for any dataset [7]. Due to this latter property, it is necessary to introduce variance in the dataset for each component learning task, in order to converge to diverse simple models, allowing different significant features of the final model to be represented in the population of simple models. This variance is obtained by: 1) perturbing the original dataset through sampling with replacement; 2) forbidding the use of some edges in the spanning arborescence, the edge blacklist being randomly generated for each component.

The paper is organized as follows. We first provide the background material through the description of required theory and previous works in section 2. Then, we describe our mixture algorithm in section 3. We finally validate in section 4 our algorithm with promising results on experiments over the challenging biological network inference application from time series data from the popular DREAM D8C1 recent challenge [8], before discussing the algorithm and its perspectives in section 5.

2 Related Works

Network inference has been studied for decades in bioinformatics context and many solutions can be grouped into the following families.

Feature selection methods break the global network inference problem into a set of N subproblems, N being the number of features in the studied biological system. The objective is to find the best neighbors of each feature in the biological model to learn, using various methods. A main strategy is to learn and rank, for each feature j, a set of (possibly linear) regression coefficients describing the function of all other features to j. Recent examples of this strategy include: the HLICORN method [9,10], where linear coefficients are computed between each possible gene and previously discovered coregulator sets; the GENIE3 [11] algorithm, where a supervised classification tree is built for each node, seen as the target, and potential parents are ranked according to their decreasing order of entropy gain between the tree's layers; BORUTA [12], a wrapper type of method around random forest classification; TIGRESS [13], which uses a *perturb and combine* strategy as this paper does. Note that this latter strategy is still significantly different from our proposal. First, TIGRESS reduces the learning problem into an independent learning problem for each node, while our proposal constrains the learned models globally, allowing here to enforce global sparsity and more generally allowing rich extensions to the method (e.g. global prior addition). Furthermore, the models to combine in this work are obtained by Lasso regression and forward selection strategy, without any quality guarantee, unlike our strategy. Finally, the approach only introduces variance in the process via data perturbation, while we propose both data and edge space sampling for each component, the latter showing significant experimental impact (cf. Section 4).

Ordinary Differential Equations (ODE) methods aim at modelling a dynamic biological system as a set of differential equations involving its features (genes, proteins, etc.). In practice, these methods can find highly accurate parameters for given equations, and have even been extended for learning the equations themselves [14], but their computational cost make them only suitable for small biological systems, which is not our target.

Pairwise score approaches aim at finding networks optimizing a sum of feature to feature scores measuring the level of correlation or dependency between them. A famous algorithm in this category is the ARACNE [15] method which computes a superset of the Chow-Liu spanning tree induced by the learning dataset, by first computing a complete graph with all pairwise mutual informations (MI), and then iteratively flagging edges for removal using a MI inequality for every triangle in the graph. Such posterior optimization, called *transitive reduction*, is indeed often very important in algorithms of this family since distinguishing between direct or indirect dependencies is not necessarily possible with the used measures.

(Dynamic) Bayesian network (DBN) learning algorithms [16] aim at finding the best factorization of the joint distribution involving the features at different consecutive time steps, as a product of conditional distributions (one per feature). Most DBN learning algorithms targetting problems of any size try to find a local optimum in the model space, through the use of a heuristic, going from one model in the space to another through iterative local perturbation of a best candidate obtained so far (best-first strategy) [17] or a set of best candidates (beam search) [18]. These algorithms often add some extra-mechanics to proceed further the first local optima, as for example random restarts or tabu search [19].

Most methods in the above mentioned families consider the input dataset as is to make network inference. In the context of small datasets, the data fragmentation issue together with the possibility of noise presence in

the dataset can misconduct the learning algorithms. In order to abstract from a single model learned from such dataset, ensemble learning strategies have been designed to implement a "wisdom of crowds" principle in a machine learning context. Historically used in a fully supervised context to average classification predictions and turn them into final majority decisions, these principles have been transposed for probability density estimation and structure learning contexts, a good example of such strategy for probabilistic models having been theorized under the Bayesian Model Averaging (BMA) [4] framework.

In BMA, the objective is to find an expected model, defined by a set of *expected features* it must satisfy, such as the dependency between its variables, the underlying graph edges or paths and so forth, assuming those features are relevant in the model space, by combining different *component* models. In practice, considering the whole space is intractable and one must rely on a subset to approximate the result, such as using Markov Chain Monte Carlo methods [20] or bagging [21].

Typical BMA methods consider the whole model space for learning which leads to the question of which heuristic to use and with which extra mechanics. This problem has been recently partially tackled in [5] but choices still remain. A recent work by Schnitzler et al. [22] considers combining more simple component models instead, namely spanning trees, and proposes extensions of the Chow-Liu algorithm [23] which show good results in their non-dynamic density estimation context. The algorithm proposed in this paper is close but differs in two main ways: it focuses at outputing an expected structure as in traditional BMA for structure learning, and thus require to make choices between the different component properties, while their work outputs a result in the form of a linear combination of each component's density probability function; it focuses on time series datasets instead of i.i.d. propositional data. Note that a preliminary adaptation of Schnitzler work for structure learning has been proposed in [24], but authors use the component models as an intermediate product to refine potential parent candidates of a greedy algorithm, while we propose an algorithm which directly infers a model from the components.

Among BMA methods, it is important to mention the seminal work in [4] about BMA structure learning through the computation of expected features using bagging on the Bayesian networks space. However, in this work, learned expected features are used as constraints for heuristic learning, which does not include component edges themselves, and whose result does not lead to significant improvement compared to the unconstrained learning counterpart. Finally, in this paper, while most work focuses on combining unconstrained Bayesian networks, few indications are also given on the results obtained by merging spanning trees, which apparently gave worse results than with unconstrained Bayesian networks. However, this can be explained by the low introduction of variance, unlike our proposal, which can lead to poor space exploration because of the components global optimality. Also, chosen experiment datasets can not be considered as being high dimensional ones and the variables/individuals ratio is less subject to data fragmentation, which can explain that more biased models perform worse. We will demonstrate in section 4 that combining arborescences in an adequate way can actually give very good results in this context.

3 Combining Spanning Arborescences for Network Inference

In this section, we propose a general algorithm for network inference from the combination of spanning arborescences. Several application oriented details are willingly not given, such as the specific scoring function used and the value of hyperparameters. These informations are detailed in Section 4 for the problem at hand.

3.1 Data representation

Let us consider a matrix representation of a dataset D consisting of n ordered time stamps (D rows) over N variables (D columns). Each column j is a sequence describing an observed variable over n time steps $\langle D_{ij} \rangle_{i \in \langle 1,\dots,n \rangle}$ and each row i describes the state of a system at time step i over the N observed variables. Our goal is to find a model of the system of interest in terms of dependencies between the observed variables at different time steps of the system. In this paper, we assume that this system is a Markov process, i.e. that each time step state only depends on the immediate previous step state, and that the transition from each state to the next state is driven by the same underlying model.

The first step for the proposed algorithm to work is to transform the $n \times N$ dataset into a $(n-1) \times 2N$ dataset D^t describing 2 consecutive time slices of the system. The transformation consists in concatenating

every pair of consecutive time steps from D into a "dynamic" example in D^t, i.e. ,we have for each row $D^t_i.$:

$$D^t_i. = [D_{i,1} \ \ldots \ D_{i,N} \ D_{i+1,1} \ \ldots \ D_{i+1,N}].$$

The given assumptions and the consequent representation allow transforming the ordered structure learning problem from D into a simpler i.i.d. structure learning problem in the dynamic examples space of D^t.

3.2 Learning component models

From a dynamic dataset D^t, the first learning step of the proposed algorithm is to compute a set of *component* models, i.e. simple models which will be combined in the second part of the algorithm.

Considering a number m of simple models to learn, we first compute m local perturbations of D^t, denoted by $\{D^{t[k]}\}_{1 \leq k \leq m}$, by sampling from D^t with replacement (bagging strategy [21]). Then, for each $D^{t[k]}$, a directed graph $G^k = (V, E^k)$, with V having one vertex for each of the N original variables, is built by first randomly choosing $\alpha \cdot N \cdot (N-1)/2$ undirected edges and then computing the two directed scores $s(A \to B)$ and $s(B \to A)$ for each of them. Finally, each graph G^k is searched for its optimum spanning arborescence A^k with respect to the score s, using the Edmonds algorithm [6].

Even if the built graphs only have one node per original variable (as opposed to two nodes, one for timestep t and one for timestep $t+1$), the semantics of an arc $X \to Y$ measures the influence of X at time t over Y at time $t+1$. This semantics is taken into account during scores' computation. Indeed, a score computation $s(X \to Y)$ is actually a score involving $D^{t[k]}_{.x}$ and $D^{t[k]}_{.(y+N)}$ columns of D^t, where x and y are the indices of variables X and Y in D. Concerning the choice of score itself, many asymmetrical scores can be used. A simple one with conditionals interpretation is the conditional entropy $H(X|Y)$. Bayesian scores, like Bayesian Dirichlet variants [7] can also be used with the advantage of being able to add prior information, relatively to each edge, to the learning task.

The edge sampling step before spanning arborescence computation is important in order to counterbalance the determinism of the Edmonds algorithm due to its global optimality. The choice of α at this step is critical and will be discussed in Section 4. On one hand, it must be low enough to avoid that the optimality of the spanning algorithm restricts the component models diversity too much before the combination step. On the other hand, it should also be high enough so that the spanning algorithm still discriminates between different models and the resulting components are not just the consequences of random sampling. The choice of sampling undirected edges instead of directly sampling directed ones is of major importance to this purpose, since it allows to obtain strongly connected G^k graphs, to guarantee the existence of a spanning arborescence for a wide range of α.

To conclude, the first learning step ensures each component to be sparse, and globally optimal in the considered arborescences space, with respect to s. Edmonds algorithm, like Kruskal one in the undirected graphs space, allows to get this optimality for each directed component in polynomial time.

3.3 Computing the composite model

Once the m component models have been learned, the second learning step aims at combining them into a composite model. In the Bayesian Model Averaging framework, this step is achieved by computing a set of *expected features* $\{\mathbf{E}(f_i)\}_i$ for the composite model, each $\mathbf{E}(f_i)$ being inferred from each component features set $\{f^k_i\}_{1 \leq k \leq m}$. In this paper, the feature space consists in the set of all possible edges in V^2, and an expected edge score is computed by counting how often that edge was present in the arborescence A^k, considering it was present in the initial weighted graph G^k. Formally, we have for all $(A, B) \in V^2$:

$$\mathbf{E}(f_{A \to B}) \approx \frac{|\,\{k \,|\, (A,B) \in \mathbf{edges}(A^k)\}\,|}{\alpha}.$$

More complex features could be considered, such as paths instead of edges or ancestor / descendant relationships, as in [4] (although the authors do not combine them in a single model). We leave these problems for future work since it would require more complex combination rules, requiring transitive reduction techniques [25], a difficult problem in the case where the input graph has cycles or is weighted.

The computation of all edges' expected scores in the composite model directly provides a ranking for those edges. A combined model is finally built from such ranking by choosing the k-best edges or all edges whose

score exceeds a given threshold. The ranking itself can be used to compare the learning results with an optimal model through an AUROC evaluation.

The Algorithm 1 summarizes the proposed approach.

Algorithm 1 The learning algorithm

Require:
D^t: a dynamic 2 slices of time learning dataset,
m: a number of component models to learn;
s: an edge directed weighting score;
α: a density for graphs setup pre-spanning arborescence;
σ: an edge weight threshold for final edges keep decision;
Ensure: a structural model from t to $t+1$ of the system
for $1 \leq k \leq m$ **do**
$D^{t[k]} := sample\_with\_replacement\_from(D^t)$
$G^k := build\_strongly\_connected\_graph(D^{t[k]}, s, \alpha)$
$A^k := edmonds\_spanning\_arborescence(G^k)$
$F^k := \mathbf{edges}(A^k)$
end for
$\forall (A,B) \in V^2 : \mathbf{E}(f_{A\to B}) := |\{k \mid (A,B) \in F^k\}| / \alpha$
return $G = choose\_top\_edges(\{\mathbf{E}(f_{A\to B})\}, \sigma)$

3.4 Complexity

Following the decomposition of the Algorithm 1, the time complexity can be expressed as the sum of two terms: one for the components computation, and another for the combination step. The components computation complexity is $m \cdot (s+g+e)$, where s (resp. g, e) is the complexity of sampling (resp. connected graph construction and Edmonds algorithm). The complexity of the sampling step is negligible here, but the construction of the graph G^k is in $\mathcal{O}(\alpha N(N-1)) \approx \mathcal{O}(N^2)$, as is the Edmonds algorithm computation with the Tarjan optimization for dense graphs [26] ($\mathcal{O}(N^2\, log\, N)$ for sparse ones). Thus the components computation part is in $\mathcal{O}(mN^2)$.

The combination part is trickier since it consists in a succession of joins between the component edgelists for further counting. Depending on the join algorithm used, this part can become the bottleneck of the overall learning approach. Indeed, a simple nested join has a time complexity in $\mathcal{O}(PQ)$ where P and Q are the number of rows in each table to join together. In our algorithm, this leads to a complexity in $\mathcal{O}(N^{2m})$. However, it is possible to considerably improve this step using better strategies, such as hash joins, running in $\mathcal{O}(P+Q)$, thus leading to a linear complexity in our settings. Note that in a purely sequential algorithm, it is not really necessary to compute joins since component edges can be counted just after arborescence computation. However, since this method is highly parallelizable due to the independence of components learning and of combinations order, it is preferable to consider this solution since the parallelization gain overcomes the joins cost in practice.

Overall, the proposed approach is the sum of a quadratic and a linear step (in a parallel configuration), and thus is of quadratic complexity.

4 DREAM 8 (HPN-DREAM) SC1B Network Inference Challenge Results

In this section, we validate our method and compare it with other algorithms through a dataset of the recent HPN-DREAM 8 Breast Cancer closed challenge [8].

4.1 Challenge and evaluation method description

The DREAM 8 SC1B subchallenge learning objective is to find the network of a synthetical biological model built using state of the art methods and biological knowledge. Simulation of this model led to the production of several time series involving 20 biological features. The data used to perform the validation of our algorithm was firstly pre-processed as described in section 3 to produce a single two-time slice dataset, the resulting data containing 80 t to $t+1$ examples over 40 temporal features.

The evaluation of learning results for this task is achieved by an official tool, the *DREAMTools* python package [27], through the computation of an AUROC score against the golden standard. In addition to com-

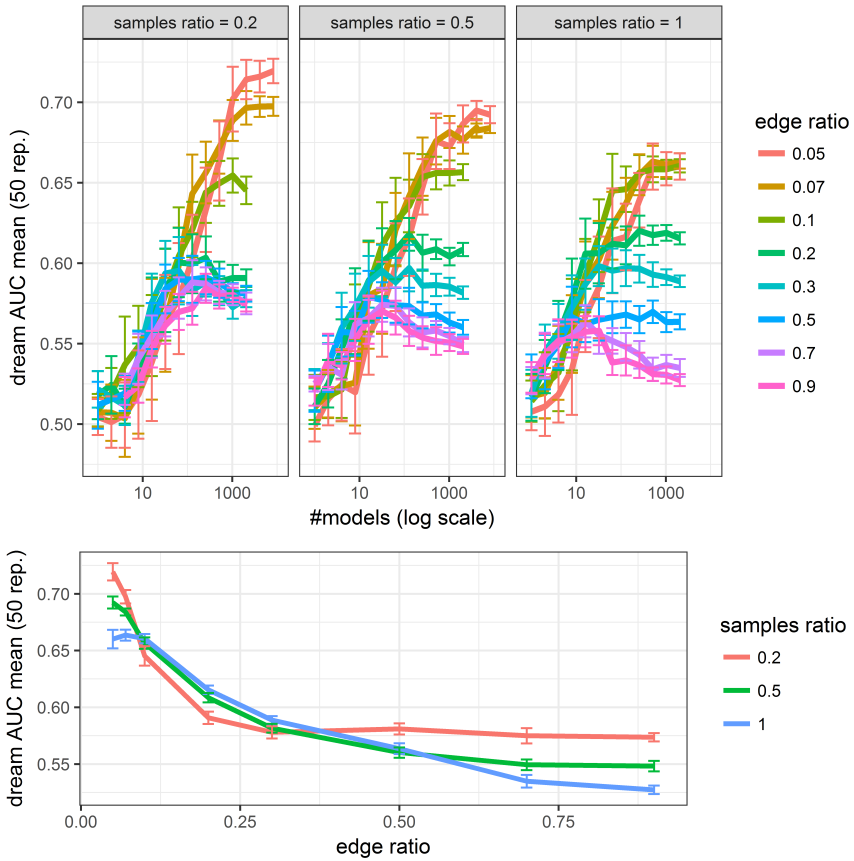


Fig. 1. (Top) mean and sds of AUC computed by DREAMTools over 50 computations as a function of the number of combined models, as well as the samples and edge ratio used for components learning. (Bottom) mean and sds of AUC computed by DREAMTools over 50 computations as a function of the edge and sample ratios. Only convergence values for increasing m are plotted.

puting scores the same way from one algorithm to another, this package also provides the expected ranking an algorithm would have reached if the challenge were still open, using all final results from the more than 100 official submissions, which allows for a cheap comparison with many algorithms of all families described in section 2.

In order to quantify the impact of several parameters on our algorithm learning quality, we have tested the method with different parametrization of the number of combined modes m, the ratio of samples n contained in each data perturbation, and the ratio of edges α present in each graph before each component learning. We used BDeu [7] gain as edge score, the difference between the $BDeu$ score of the $A \rightarrow B$ local structure and the no edge one. Namely for an edge $A \rightarrow B$: $BDeu(parents(B) = \{A\}) - BDeu(parents(B) = \emptyset)$.

4.2 Results

Results for many parametrizations, given in Figure 1(Top), show different clear trends. Firstly, we can see that for small edge ratios, the obtained AUC seems to monotonically increase with the number of combined models, until reaching plateaus. For bigger ratios, the trend seems to be mostly observable, but the higher the

sampling ratio, the lower the minimum edge ratio needs to be to show this trend. Additionally, we can observe that the convergence AUC value tends to increase whenever any of the edge or the sample ratio decrease, which is clearer in Figure 1(Bottom). These results seem to indicate that focusing on smaller parts of the available information for each component, while aggregating a higher number of them for final consensus, seem to give the best results, which confirms the requirement for components diversity in order to give a good consensus. Extra experiments done and not displayed here show that smaller edge and sample ratios break the observed trends. For edge ratio, this is predictable, since the minimum value displayed of 0.05 corresponds to an average number of 2 neighbors per node (considering we add the reverse edges for each sampled edge, to ensure we can compute a spanning arborescence), which is the minimal number of neighbors required for the algorithm to make a choice. Lower values actually lead the spanning arborescence algorithm to just select most available edges in the graphs it is given. For samples ratio, it seems to indicate a limit from which the dataset is too small to capture faithful enough information.

Concerning the expected ranking for the different results, our approach is very promising since it reached the 3^{rd} position for the best mean AUC obtained over the different parametrizations, outperforming GENIE3, ARACNE, all heuristic oriented Bayesian network methods, as well as all linear and most non-linear regression methods, all ODE and all ensemble learning solutions.

4.3 Discussion

At the moment, the gap with the best performance is of 0.045. A particularity of the considered DREAM subchallenge is that 3 out of the 20 biological features are actually fake nodes, supposed to have no correlation with the others. This shows a limitation of our approach in its current form: learning spanning trees means that every node will get one parent per component, even if there is no true correlation. Note that this problem is not necessarily easy to solve, since there is also a tendency for such spurious correlations to be non-uniformly distributed. Indeed, the optimization of the spanning arborescence score encourages to keep the apparently more correlated pairs of nodes, so the ones with the most biased noise are chosen. Since sample and edge samplings are uniformly done, there is a high probability for a restricted number of parents to appear in each component for a fake node. In practice, this means that a simple pruning of the components is not enough. Future work will address this issue.

5 Conclusion

In this paper, we have presented a network inference learning algorithm based on the combination of multiple spanning arborescences learned over multiple perturbation of the original dataset, with enforced diversity through edge sampling, showing promising results in practice on a recent DREAM challenge.

Experiments have more particularly shown that combining more models together with more diversity, involving a decrease of both sample and edge ratios in the currently defined parameter space, leads to better convergence values. It is encouraged to use this strategy in a quite extreme way, since best performances obtained in the experiments are achieved by situations where both ratios are very low. The only warning would be to still allow the Edmonds algorithm to have choice, in order not to make the components completely random. We have also seen in section 4 the impact of fake nodes on the results, and the difficulty of identifying them whenever the spanning arborescence assign most nodes a parent. This issue has a significant impact on the current results since removing edges involving fake nodes would lead to a top position of the approach.

Future works will address the limitations of the current algorithm, such as its sensitivity to fake nodes. More advanced extensions will also be investigated, such as the introduction of priors, really important in biological contexts, modular capabilities, which is becoming a standard in recent methods to abstract from a model complexity, and different component combination rules to preserve extra properties in the consensus model, such as paths or path lengths.

References

[1] Michael Hecker, Sandro Lambeck, Susanne Toepfer, Eugene van Someren, and Reinhard Guthke. Gene regulatory network inference: Data integration in dynamic models—a review. *Biosystems*, 96(1):86 – 103, 2009.

[2] Karen Sachs, Omar Perez, Dana Pe'er, Douglas A. Lauffenburger, and Garry P. Nolan. Causal protein-signaling networks derived from multiparameter single-cell data. *Science*, 308(5721):523–529, 2005.

[3] Ludwig Geistlinger, Gergely Csaba, Simon Dirmeier, Robert Küffner, and Ralf Zimmer. A comprehensive gene regulatory network for the diauxic shift in saccharomyces cerevisiae. *Nucleic acids research*, page gkt631, 2013.

[4] Nir Friedman, Moises Goldszmidt, and Abraham Wyner. Data analysis with bayesian networks: A bootstrap approach. In *Proceedings of the Fifteenth conference on Uncertainty in artificial intelligence*, pages 196–205. Morgan Kaufmann Publishers Inc., 1999.

[5] Bradley M Broom, Kim-Anh Do, and Devika Subramanian. Model averaging strategies for structure learning in bayesian networks with limited data. *BMC bioinformatics*, 13(13):S10, 2012.

[6] Jack Edmonds. Optimum branchings. *Mathematics and the Decision Sciences, Part*, 1:335–345, 1968.

[7] David Heckerman, Dan Geiger, and David M Chickering. Learning bayesian networks: The combination of knowledge and statistical data. *Machine learning*, 20(3):197–243, 1995.

[8] Steven M Hill, Laura M Heiser, Thomas Cokelaer, Michael Unger, Nicole K Nesser, Daniel E Carlin, Yang Zhang, Artem Sokolov, Evan O Paull, Chris K Wong, et al. Inferring causal molecular networks: empirical assessment through a community-based effort. *Nature methods*, 13(4):310–318, 2016.

[9] I Chebil, Rémy Nicolle, G Santini, Céline Rouveirol, and Mohamed Elati. Hybrid method inference for the construction of cooperative regulatory network in human. *IEEE transactions on nanobioscience*, 13(2):97–103, 2014.

[10] Rémy Nicolle, François Radvanyi, and Mohamed Elati. Coregnet: reconstruction and integrated analysis of co-regulatory networks. *Bioinformatics*, page btv305, 2015.

[11] Va Anh Huynh-Thu, Alexandre Irrthum, Louis Wehenkel, and Pierre Geurts. Inferring regulatory networks from expression data using tree-based methods. *PLOS ONE*, 5(9):1–10, 09 2010.

[12] Miron B Kursa and Witold R Rudnicki. Feature selection with the boruta package. *Journal of Statistical Software*, 36(11):1–13, 2010.

[13] Anne-Claire Haury, Fantine Mordelet, Paola Vera-Licona, and Jean-Philippe Vert. Tigress: Trustful inference of gene regulation using stability selection. *BMC Systems Biology*, 6(1):145, 2012.

[14] Alex Greenfield, Christoph Hafemeister, and Richard Bonneau. Robust data-driven incorporation of prior knowledge into the inference of dynamic regulatory networks. *Bioinformatics*, 29(8):1060–1067, 2013.

[15] Adam A Margolin, Ilya Nemenman, Katia Basso, Chris Wiggins, Gustavo Stolovitzky, Riccardo D Favera, and Andrea Califano. Aracne: an algorithm for the reconstruction of gene regulatory networks in a mammalian cellular context. *BMC bioinformatics*, 7(Suppl 1):S7, 2006.

[16] Min Zou and Suzanne D Conzen. A new dynamic bayesian network (dbn) approach for identifying gene regulatory networks from time course microarray data. *Bioinformatics*, 21(1):71–79, 2005.

[17] Judea Pearl. Heuristics: intelligent search strategies for computer problem solving. 1984.

[18] Peter Norvig. *Paradigms of artificial intelligence programming: case studies in Common LISP*. Morgan Kaufmann, 1992.

[19] Jan Karel Lenstra. *Local search in combinatorial optimization*. Princeton University Press, 2003.

[20] Walter R Gilks. Markov chain monte carlo. *Encyclopedia of Biostatistics*, 2005.

[21] Leo Breiman. Bagging predictors. *Machine learning*, 24(2):123–140, 1996.

[22] François Schnitzler, Sourour Ammar, Philippe Leray, Pierre Geurts, and Louis Wehenkel. *Efficiently Approximating Markov Tree Bagging for High-Dimensional Density Estimation*, pages 113–128. Springer Berlin Heidelberg, Berlin, Heidelberg, 2011.

[23] C Chow and C Liu. Approximating discrete probability distributions with dependence trees. *IEEE transactions on Information Theory*, 14(3):462–467, 1968.

[24] Sourour Ammar and Philippe Leray. *Mixture of Markov Trees for Bayesian Network Structure Learning with Small Datasets in High Dimensional Space*, pages 229–238. Springer Berlin Heidelberg, Berlin, Heidelberg, 2011.

[25] Andrea Pinna, Sandra Heise, Robert J Flassig, Alberto De La Fuente, and Steffen Klamt. Reconstruction of large-scale regulatory networks based on perturbation graphs and transitive reduction: improved methods and their evaluation. *BMC systems biology*, 7(1):73, 2013.

[26] Robert Endre Tarjan. Finding optimum branchings. *Networks*, 7(1):25–35, 1977.

[27] Thomas Cokelaer, Mukesh Bansal, Christopher Bare, Erhan Bilal, Brian M Bot, Elias Chaibub Neto, Federica Eduati, Alberto de la Fuente, Mehmet Gönen, Steven M Hill, et al. Dreamtools: a python package for scoring collaborative challenges. *F1000Research*, 4, 2015.

Long-term Tracking of Budding Yeast Cells with CellStar

Cristian VERSARI[1], Szymon STOMA[2], Kirill BATMANOV[1], Artémis LLAMOSI[3,4], Filip MROZ[5], Adam KACZMAREK[5], Matt DEYELL[3], Cédric LHOUSSAINE[1], Pascal HERSEN[3] and Gregory BATT[4]

[1] CRIStAL – Univ. Lille 1, bât M3 ext. Av. Carl Gauss, 59655, Villeneuve d'Ascq, France
[2] ScopeM - ETH Zurich, Wolfgang-Pauli-Str. 14, 8093, Zurich, Switzerland
[3] Lab. MSC, bât. Condorcet 10 rue A. Domon et L. Duquet, 75205, Paris cedex 13, France
[4] Inria Saclay, Bât A. Turing 1 rue H. D'Estienne d'Orves, 91120, Palaiseau, France
[5] Inst. of Computer Science – Univ. of Wroclaw, Joliot-Curie 15, 50-383, Wroclaw, Poland

Corresponding Author: cristian.versari@univ-lille1.fr

Abstract

Observing cellular processes at the single cell level is often necessary to understand how cells respond to endogenous and environmental changes. Used in combination with fluorescence reporter techniques, flow cytometry and time-lapse microscopy are arguably the two most widely employed quantitative single-cell observation approaches. The former provides great statistical details on the diversity of the studied cell population, whereas the latter provides longitudinal information on single cells: individual cells can be tracked in time. This is a decisive advantage to investigate a number of important biological problems, including chronological aging, epigenetic heritability, and dynamic features such as cell-cycle and circadian oscillations in non-synchronized cell populations.

However, the capability to extract single cell traces from microscopy images in a fully-automated manner is a necessary prerequisite to obtain conclusions that are valid and biologically relevant in long lasting experiments. Incorrect assignments (e.g. two cells exchanged at some time point) can possibly hide interesting features, or worse, create spurious information. Although such incorrect assignments are expected to be relatively rare at each time point, a simple analysis shows that the number of correct traces decreases rapidly with the duration of the experiment: even for yeast cells that have relatively regular shapes, no solution has been proposed that reaches the high quality required for long-term experiments for segmentation and tracking based on brightfield images.

In this demo we present CellStar [1], a tool chain designed to achieve good performance in long-term experiments. The key features are the use of a new variant of parametrized active rays for segmentation, a neighbourhood-preserving criterion for tracking, and the use of an iterative approach that incrementally improves segmentation and tracking quality. A graphical user interface enables manual corrections of segmentation and tracking errors and their use for the automated correction of other, related errors and for parameter learning.

In [1] we created a benchmark dataset with manually analysed images and compared CellStar with six other tools, showing its high performance, notably in long-term tracking. As a community effort, we set up a website, the Yeast Image Toolkit, with the benchmark and the Evaluation Platform to gather this and additional information provided by others.

References

[1] Cristian Versari, Szymon Stoma, Kirill Batmanov, Artémis Llamosi, Filip Mroz, Adam Kaczmarek, Matt Deyell, Cédric Lhoussaine, Pascal Hersen and Gregory Batt. Long-term tracking of budding yeast cells in brightfield microscopy: CellStar and the Evaluation Platform. *Journal of The Royal Society Interface*, 14(127), 2016.

JOBIM 2017 Highlight

Listeriomics: A Multi-Omics Interactive Web Platform for Systems Biology of the Model Pathogen *Listeria*

Christophe Bécavin[1,2,3,4], Mikael Koutero[1,2,3], Nicolas Tchitchek[5], Franck Cerutti[6], Pierre Lechat[4], Nicolas Maillet[4], Claire Hoede[6], Hélène Chiapello[6], Christine Gaspin[6], and Pascale Cossart[1,2,3]

[1] Institut Pasteur, Unité des Interactions Bactéries-Cellules, Département de Biologie Cellulaire et Infection, F-75015 Paris, France

[2] INSERM, U604, F-75015 Paris, France

[3] INRA, USC2020, F-75015 Paris, France

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

[5] CEA, Division of Immuno-Virology, Institute of Emerging Diseases and Innovative Therapies, IDMIT Center, Fontenay-aux-Roses 92265, France

[6] MIAT, Université de Toulouse, INRA, Castanet-Tolosan, France

Corresponding Author: christophe.becavin@pasteur.fr

Paper Reference: **Bécavin et al. (2016) Listeriomics: An Interactive Web Platform for Systems Biology of *Listeria*. mSystems. http://dx.doi.org/10.1128/mSystems.00186-16**

Abstract*: As for many model organisms, the amount of Listeria omics data produced has recently increased exponentially. There are now >80 published complete Listeria genomes, around 350 different transcriptomic data sets, and 25 proteomic data sets available. The analysis of these data sets through a systems biology approach and the generation of tools for biologists to browse these various data are a challenge for bioinformaticians. We have developed a web-based platform, named Listeriomics, that integrates different tools for omics data analyses, i.e., (i) an interactive genome viewer to display gene expression arrays, tiling arrays, and sequencing data sets along with proteomics and genomics data sets; (ii) an expression and protein atlas that connects every gene, small RNA, antisense RNA, or protein with the most relevant omics data; (iii) a specific tool for exploring protein conservation through the Listeria phylogenomic tree; and (iv) a coexpression network tool for the discovery of potential new regulations. Our platform integrates all the complete Listeria species genomes, transcriptomes, and proteomes published to date. This website allows navigation among all these data sets with enriched metadata in a user-friendly format and can be used as a central database for systems biology analysis.*
Link: http://listeriomics.pasteur.fr/

Keywords Listeria, genomics, transcriptomics, proteomics, systems biology

Summary

Listeria monocytogenes is a foodborne pathogen responsible for foodborne infections with a mortality rate of 25%. This pathogen is responsible for gastroenteritis, sepsis, and meningitis and can cross three host barriers, the intestinal, placental, and blood-brain barriers [1]. Over the past three decades *Listeria has become a model organism for host-pathogen interactions*, leading to critical discoveries in a broad range of fields [2], including virulence-factor regulation, cell biology, and bacterial pathophysiology. To study these mechanisms, several genomics, transcriptomics, and proteomics data sets have been produced.

The analysis of these data sets through a systems biology approach and *the generation of tools for biologists to browse these various data are a challenge for bioinformaticians*. We have developed a web-based platform, named Listeriomics (http://listeriomics.pasteur.fr/), that integrates different tools (see Figure 1) for omics data analyses, i.e., (i) *an interactive genome viewer* to display gene expression arrays, tiling arrays, and

sequencing data sets along with proteomics and genomics data sets; (ii) *an expression and protein atlas* that connects every gene, small RNA, antisense RNA, or protein with the most relevant omics data; (iii) a specific tool for exploring *protein conservation through the Listeria phylogenomic tree*; and (iv) a *coexpression network tool* for the discovery of potential new regulations.

To our knowledge, none of the referent databases dedicated to model organisms such as *E. coli* or *B. subtilis* integrates as many data sets and visualization tools as the Listeriomics resource does. User experience and feedback from our collaborators using the Listeriomics interface [3] for the past 5 years were driving forces in organizing and improving the way to access data and tools. Our main purpose was to *design an easy-to-use website with a dynamic interface for biologists* wanting to access the different heterogeneous "omics" data sets available for Listeria. *As such our website should interest the JOBIM community as it shows an example of extensive multi-omics data integration for model organism*.

The website is developed with an in-house platform which we named BACNET. It is written in Java language, with BioJava and Eclipse RAP APIs. We designed the platform so bioinformaticians wishing to *create they own multi-omics website for another organism can do so with few efforts*. *We believed the description of the BACNET platform will also interest the JOBIM community*.

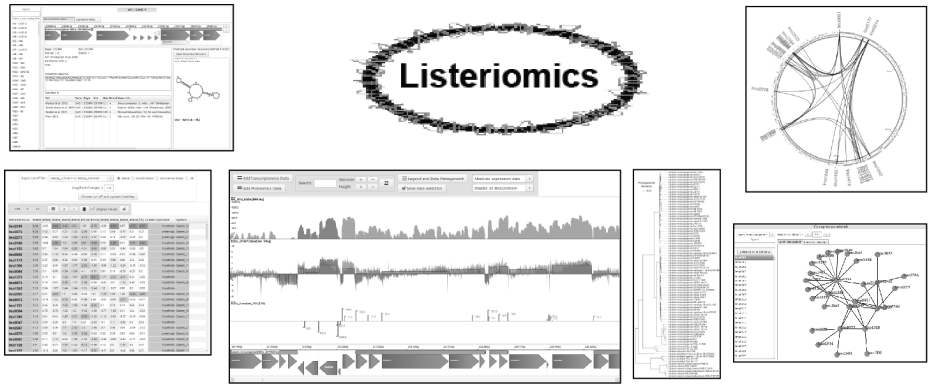


Figure 1: Overview of the different tools available in Listeriomics website

References

[1] P. Cossart, "Illuminating the landscape of host-pathogen interactions with the bacterium Listeria monocytogenes.," *Proc. Natl. Acad. Sci. U. S. A.*, Nov. 2011.

[2] A. Lebreton, F. Stavru, S. Brisse, and P. Cossart, "1926–2016: 90 Years of listeriology," *Microbes Infect.*, vol. 18, no. 12, pp. 711–723, Dec. 2016.

[3] F. Impens, N. Rolhion, L. Radoshevich, C. Bécavin, M. Duval, J. Mellin, F. García del Portillo, M. G. Pucciarelli, A. H. Williams, and P. Cossart, "N-terminomics identifies Prli42 as a membrane miniprotein conserved in Firmicutes and critical for stressosome activation in Listeria monocytogenes," *Nat. Microbiol.*, vol. 2, no. February, p. 17005, 2017.

AskOmics, a web tool to integrate and query biological data using semantic web technologies

Xavier GARNIER[1,2], Anthony BRETAUDEAU[1,3], Olivier FILANGI[1,3], Fabrice LEGEAI[1,4], Anne SIEGEL[2] and Olivier DAMERON[2]

[1] 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

[2] DYLISS (INRIA - IRISA) - INRIA, Université de Rennes 1, CNRS : UMR6074 - Campus de Beaulieu, F-35 042 Rennes Cedex, France

[3] Plateforme bioinformatique GenOuest - Université de Rennes 1, Biogenouest - France

[4] GENSCALE (INRIA - IRISA) - Université de Rennes 1, CNRS : UMR6074, INRIA - Campus de Beaulieu, F-35 042 Rennes Cedex, France

Corresponding author: xavier.garnier@irisa.fr

Research programs involving genetics, genomics and epigenetics are quickly growing. Lot of experiments producing large amount of data are now feasible in the frame of a laboratory. As well, the tools analyzing the data generated by these experiments are often available. Results of these experiments can be stored and structured into files and loaded into databases, and lots of these databases are public. However, these bases are usually implemented according to schemes and techniques that do not allow their interoperability in an easy manner.

The technologies from the Semantic Web, especially RDF and SPARQL are one of the key elements for combining databases, which has led to the emergence of linked data. It is based on triples (subject, predicate and objects) describing the relationships between elements stored into interoperable triplestores allowing distributed querying. Because of its flexibility, versatility and ontology-awareness, numerous biological databases, such as UniProt, PubChem or ChEBI and Reactome at EBI, give access to their data via a SPARQL endpoint.

We present AskOmics, a new software, that uses the Semantic web technologies, which helps to integrate multiple format of data and query them through a user-friendly interface. AskOmics is a free and open-source software (AGPL licence) available on GitHub (https://github.com/askomics/askomics).

AskOmics supports both intuitive data integration and querying while shielding a non-expert user from most of the technical difficulties underlying the web semantic technologies. Because large and heterogeneous biological datasets are often difficult to integrate, AskOmics users can provide simple tabulation-separated files (TSV), that are transformed automatically into RDF triples, then stored into a triplestore. Finally, for data querying, AskOmics provides a visually intuitive interface to obtain a comprehensive view of the biological study.

During data integration, user provides input files in common formats (currently TSV and GFF) to be converted into RDF triples. AskOmics generates triples corresponding to the data (the content), and also triples which describe the data (the abstraction). Triples are loaded in a triplestore in order to persist data and optimize queries.

The query interface is composed of a dynamic graph at the left and a right view for filtering attributes. On the graph, each node represents an entity. Entities are linked between them with arrows. Attributes of the selected entities are displayed on the right view. To build the graph, AskOmics query the abstraction. Users build their queries by starting from a node of interest and sequentially select its neighbors and filter on attributes, creating a path on the abstraction. This path is converted into a SPARQL query and sent to the triplestore. Finally, the results are displayed as a table and can be downloaded as a TSV file.

AskOmics has been applied successfully to the analysis of large scale datasets on the aphid embryogenesis, on the variablility of Brassicaceae in response to clubroot disease, and to the analysis of biological pathways.

Regulatory and signaling network assembly through Linked Open Data

Marie LEFEBVRE[1], Jérémie BOURDON[2], Carito GUZIOLOWSKI[2] and Alban GAIGNARD[1]
[1] Nantes Academic Hospital, CHU de Nantes, France
[2] LS2N - UMR 6004, University of Nantes, Ecole Centrale de Nantes, France

Corresponding author: marie.lefebvre@univ-nantes.fr, alban.gaignard@univ-nantes.fr

1 Introduction and problem statement

Nowadays, huge efforts address the organization of biological knowledge through linked open databases. These databases can be automatically queried to reconstruct a large variety of biological networks such as regulatory or signaling networks. Assembling networks still implies manual operations due to *(i)* source-specific identification of biological entities, *(ii)* source-specific semantics for entity-entity relationships, *(iii)* proliferating heterogeneous life-science databases with redundant information and *(iv)* the difficulty of recovering the logical flow of a biological pathway due to the bidirectionality of chemical reactions. Homogenization of biological networks is therefore costly and error-prone. Existing tools such as the ReactomeFIPlugIn of Cytoscape 3.0[1] or STRING[2] allow to link entities to each other or to identify an entity's membership to a single pathway. Nevertheless, they are still limited in the global modeling aspects (logical rules inferred from the knowledge representation). Here, we present a framework to automate the assembly of regulatory and signaling networks in the context of tumor cells modeling.

2 Approach

Our framework is based on Semantic Web technologies. It addresses *(i)* the uniform identification of multi-source biological entities, *(ii)* the description of labeled directed graphs through RDF, and *(iii)* the use of BioPAX [3] as a semantic reference. We consider a list of target gene names or IDs as entry points. The first step consists in retrieving transcription factors (TFs) controlling these target genes. Then, the second step consists in considering the TFs as new entry points for the reconstruction algorithm. The full regulatory network is finally assembled by iteratively applying the second step until no new TFs can be found.

3 Demonstration

To assemble networks our algorithm queries PathwayCommons[4] through its SPARQL endpoint and retrieves a graph of TFs associated to target genes. We developed a web tool that displays the biological network assembly step by step, allowing users to interact with the reconstruction process and to visually shape the network. Through this web tool, it is also possible to launch a command line tool (Java) to address larger scale input gene lists. These tools have been deployed on the BiRD Cloud infrastructure. From a list of 1800 targets genes, we were able to assemble in less than 3 minutes a graph of 1474 nodes and 12303 edges.

4 Discussion and Conclusion

As future works, we aim at integrating drug-target informations (*e.g.* KEGGdrug, DrugBank) through SPARQL federated queries to get insights on *(i)* tumor cells growth and *(ii)* drug response on patient and cell lines gene expression data. Our tool is freely available at https://github.com/symetric-group/bionets-demo

Acknowledgements

This work was supported by the BiRD bioinformatics facility, the SyMeTRIC project and the GRIOTE project.

References

[1] W., Guanming et al. A human functional protein interaction network and its application to cancer data analysis. *Genome Biology*, 11(5):R53, 2010.

[2] D., Szklarczyk et al. STRING v10: protein-protein interaction networks, integrated over the tree of life. *Nucleic Acids Res.*, 43:D447–D452, 2015.

[3] E., Demir et al. The biopax community standard for pathway data sharing. *Nature Biotechnology*, 28:935–942, 2010.

[4] E.G., Cerami et al. Pathway commons, a web resource for biological pathway data. *Nucleic Acids Res.*, 39:D685–90, 2011.

Scientific Workflows for Computational Reproducibility in the Life Sciences: Status, Challenges and Opportunities

Sarah COHEN-BOULAKIA[1,2,3], Khalid BELHAJJAME[4], Olivier COLLIN[5], Jérôme CHOPARD[6], Christine FROIDEVAUX[1], Alban GAIGNARD[7], Konrad HINSEN[8], Pierre LARMANDE[9], Yvan LE BRAS[10], Frédéric LEMOINE[11], Fabien MAREUIL[12,13], Hervé MÉNAGER[12,13], Christophe PRADAL[14,2] and Christophe BLANCHET[15]

[1] LRI, Univ. Paris-Sud, CNRS UMR 8623, Univ. Paris-Saclay
[2] Inria, VirtualPlants, Montpellier
[3] Inria, Zenith, Montpellier
[4] Lamsade, Univ. Paris-Dauphine, CNRS UMR 7243, , Paris
[5] IRISA, Rennes, France
[6] INRA, UMR729, MISTEA, Montpellier
[7] Nantes Academic Hospital, CHU de Nantes
[8] Centre de Biophysique Moléculaire, CNRS UPR4301, Orléans
[9] IRD, DIADE, F-34394 Montpellier
[10] EnginesOn / INRIA, Rennes
[11] Institut Pasteur, Unité Bioinformatique Evolutive, C3BI, USR 3756 IP CNRS, Paris
[12] Institut Pasteur, Hub Bioinformatique et Biostatistique, C3BI, USR 3756 IP CNRS, Paris
[13] Institut Pasteur, Centre d'Informatique pour la Biologie, DSI, Paris
[14] CIRAD, UMR AGAP, Montpellier
[15] Institut Français de Bioinformatique, IFB-core, CNRS, UMS 3601, Gif-sur-Yvette

Corresponding author: cohen@lri.fr

Reference paper: **Cohen-Boulakia *et al.* (2017) Scientific Workflows for Computational Reproducibility in the Life Sciences: Status, Challenges and Opportunities. *Future Generation Computer Systems*.**
http://dx.doi.org/10.1016/j.future.2017.01.012

Abstract *With the development of new experimental technologies, an avalanche of data has to be computationally analyzed for scientific advancements and discoveries to emerge. Faced with the complexity of analysis pipelines, the large number of computational tools, and the enormous amount of data to manage, there is compelling evidence that many if not most scientific discoveries will not stand the test of time: increasing the reproducibility of computed results is of paramount importance. The objective we set out in this paper is to place scientific workflows in the context of reproducibility: We define several levels of reproducibility; we characterize and define the criteria that need to be catered for by reproducibility-friendly scientific workflow systems; we use such criteria to place several representative and widely used workflow systems and companion tools within such a framework; we discuss the remaining challenges posed by reproducible scientific workflows in the life sciences. Our study was guided by three use cases from the French community, involving in silico experiments.*

Keywords Reproducibility;Scientific Workflows;Provenance; Packaging environments.

Novel technologies have led to the generation of very large volumes of data at an unprecedented rate. This is particularly true for the life sciences, where, for instance, innovations in Next Generation Sequencing (NGS) have led to a revolution in genome sequencing. Current instruments can sequence several hundreds of human genomes in one week whereas more than ten years have been necessary for the first human genome. Many laboratories have thus acquired NGS machines, resulting in an avalanche of data which has to be further analyzed using a series of tools and programs for new scientific knowledge and discoveries to emanate. The same kind of situation occurs in completely different biological domains, such as plant phenotyping which aims at understanding the complexity of interactions between plants and environments in order to accelerate the discovery of new genes and traits thus optimize the use of genetic diversity under different environments. Here, thousands of plants are grown in controlled environments, capturing a lot of information and generating huge amounts of raw data to be stored and then analyzed by very complex computational analysis pipelines for scientific advancements and discoveries to emerge.

Faced with the complexity of analysis pipelines designed, the number of computational tools available and the amount of data to manage, there is compelling evidence that the large majority of scientific discoveries will not stand the test of time: increasing reproducibility of results is of paramount importance.

Over the recent years, many authors have drawn attention to the rise of purely computational experiments which are not reproducible (see in particular [1]). Major reproducibility issues have been highlighted in a very large number of cases: while [2] has shown that even when very specific tools were used, textual description of the methodology followed was not sufficient to repeat experiments, [3] has focused on top impact factor papers and shown that insufficient data were made available by the authors to make experiments reproducible, despite the data publication policies recently put in place by publishers.

Scientific communities in different domains have started to act in an attempt to address this problem. Prestigious conferences (such as two major conferences from the database community, namely, VLDB [15] and SIGMOD [16]) and journals such as PNAS [17], Biostatistics, Nature, and Science, to name only a few, encourage or require published results to be accompanied by all the information necessary to reproduce them. However, making their results reproducible remains a very difficult and extremely time-consuming task for most authors. In the meantime, considerable efforts have been put into the development of *scientific workflow management systems*. They aim at supporting scientists in developing, running, and monitoring chains of data analysis programs. A variety of systems (Galaxy, OpenAlea,...) have reached a level of maturity that allows them to be used by scientists for their bioinformatics experiments, including analysis of NGS or plant phenotyping data. By capturing the exact methodology followed by scientists (in terms of experimental steps associated with tools used) scientific workflows play a major role in the reproducibility of experiments.

The propose of this paper is thus to better understand the core problematic of reproducibility in the specific context of scientific workflow systems. We aim to provide answers to the following key points: How can we define the different levels of reproducibility that can be achieved when a workflow is used to implement an *in silico* experiment? Which are the criteria of scientific workflow systems that make them *reproducibility-friendly*? What is concretely offered by the scientific workflow systems in use in the life science community to deal with reproducibility? Which are the open problems to be tackled in computer science (in algorithmics, systems, knowledge representation etc.) which may have huge impact on the problems of reproducing experiments when using scientific workflow systems?

Accordingly, we make the following five contributions: We present three real use cases involving in silico experiments, and elicit concrete reproducibility issues that they raise. We define several kinds of reproducibility that can be reached when scientific workflows are used to perform experiments. We characterize and define the criteria that need to be catered for by *reproducibility-friendly* scientific workflow systems. Using the framework of the criteria identified, we place several representative and widely used workflow systems and companion tools within such a framework. We go on to discuss the challenges posed by reproducible scientific workflows in the life sciences and describe the remaining opportunities of research in several areas of computer science which may address them.

This paper is the result of a large collaborative work between several members of the French Bioinformatics community. Ongoing work includes organizing hackathons to concretely test the ability of the various workflow systems to deal with reproducibility.

Acknowledgements

The authors acknowledge the support of GDR CNRS MaDICS, programme CPER Région Bretagne "CeSGO", and programme Région Pays de la Loire "Connect Talent" (SyMeTRIC). We acknowledge funding by the call "Infrastructures in Biology and Health" in the framework of the French "Investments for the Future" (ANR-11-INBS-0012 and ANR-11-INBS-0013).

References

[1] Victoria Stodden, Peixuan Guo, and Zhaokun Ma. Toward reproducible computational research: an empirical analysis of data and code policy adoption by journals. *PloS one*, 8(6):e67111, 2013.

[2] Anton Nekrutenko and James Taylor. Next-generation sequencing data interpretation: enhancing reproducibility and accessibility. *Nature Reviews Genetics*, 13(9):667–672, 2012.

[3] Alawi A Alsheikh-Ali, Waqas Qureshi, Mouaz H Al-Mallah, and John PA Ioannidis. Public availability of published research data in high-impact journals. *PloS one*, 6(9):e24357, 2011.

15. International conference on Very Large Data Bases
16. ACM's Special Interest Group on Management Of Data.
17. http://www.pnas.org/site/authors/format.xhtml

Sequanix: A Dynamic Graphical Interface for Snakemake Workflows

Dimitri DESVILLECHABROL[1], Rachel LEGENDRE[1,2], Christiane BOUCHIER[1], Sean KENNEDY[1] and Thomas COKELAER[1,2]

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

Corresponding author: dimitri.desvillechabrol@pasteur.fr, thomas.cokelaer@pasteur.fr

Sequencing platforms produce large and heterogeneous data sets. For example, the Biomics pole (Institut Pasteur, CITECH) provides sequencing data related to transcriptomics, genomics and metagenomics. Before delivering results, platforms need to perform quality controls. Nevertheless, these platforms may also provide dedicated analysis (e.g. variant caller, de-novo).

There are many sequencing pipelines available. We provide some of them in the Sequana project. The particularity of the pipelines available in Sequana is that they are based on the Snakemake framework. Snakemake pipelines are written in Python with a rule-based syntax; a configuration file is also required (YAML format) [1]. The Snakemake pipelines targets an audience of developers since they require the pipelines to be run on cluster with command line arguments; configuration file also need to be edited.

Sequana project is made of (i) a Python library, (ii) a set of Snakemake pipelines and (iii) standalones [2]. In order to expose the Snakemake pipelines to all kind of users, we develop Sequanix, a standalone Graphical User Interface (GUI) based on Python and PyQt. Sequanix can launch Snakemake pipelines available in Sequana without the need for command line interface. Indeed, the GUI can be used on a cluster (with a display). So the knowledge of editor such as VIM is not necessary anymore. Therefore, Sequanix fills the gap between Snakemake developers and their end-users.

Pipelines provided by Sequana are directly available in a dropdown box. When a pipeline is chosen, the embedded configuration file is automatically loaded as a form. This form can be edited without the need for a text editor. It also have convenient features: if a parameter name ends in _file or _browser, a dedicated widget (file or directory browser) is shown instead of a simple line edit; if comments are available in the configuration file, we interpret and show them as tooltips in the GUI. Once a pipeline and a working directory are set, the project can be saved, the workflow visualised as a directed acyclic graph (DAG) and finally the pipeline can be executed. The interface displays the Snakemake standard output and the progress.

Sequanix exposes all Sequana pipelines (snakemake-based) within a graphical interface. Yet, many Snakemake pipelines are developed by a large community of developers especially in the NGS field. So, we extended the GUI so that any external Snakemake pipeline can also be imported and executed through the interface making Sequanix a generic tool for any Snakemake pipeline.

References

[1] Johannes Köster and Sven Rahmann. Snakemake - a scalable bioinformatics workflow engine. *Bioinformatics*, 28(19):2520-2522, 2012.

[2] Sequana: a set of flexible genomic pipelines for processing and reporting NGS analysis (http://github.com/sequana/sequana). Documentation available on http://sequana.readthedocs.io

Biosphère : un portail web haut niveau pour une utilisation bioinformatique des clouds

Bryan BRANCOTTE[1], Mohamed BEDRI[1], Jonathan LORENZO[1], Sandrine PERRIN[1], Frédéric SÉNÉ[1], Awa SEPOU NGAÏLO[1], Christophe BLANCHET[1], Jean-François GIBRAT[1]

[1] CNRS, UMS 3601 ; Institut Français de Bioinformatique, IFB-core, Avenue de la Terrasse, F-91190 Gif-sur-Yvette, France
Corresponding Author: bryan.brancotte@france-bioinformatique.fr

Le cloud, par ses promesses de machines à grande mémoire et clusters de calcul conséquents, se pose comme un sérieux candidat pour répondre aux problématiques tant scientifiques que techniques amenées par le déluge de données en bioinformatique. Afin de s'appuyer sur le cloud, il faut rendre son utilisation simple, intuitive, et conserver un haut niveau de sécurité. C'est pour répondre à ce défi technique, et dans le cadre du projet CYCLONE, que nous avons construit le portail web Biosphère (https://biosphere.france-bioinformatique.fr).

Le projet CYCLONE (http://www.cyclone-project.eu) est un projet Horizon-2020 « Actions d'innovation » financé par la Commission Européenne (projet 644925). Il vise à répondre aux challenges posés par des applications, dont certaines en bioinformatique, en intégrant et améliorant des solutions open-source pour la gestion de clouds. Les objectifs sont une gestion unifiée de différents clouds, le déploiement et le maintien de plateformes de traitement de données complexes dans une architecture multi-cloud, une réactivité et une élasticité dans l'utilisation des ressources proposées. Dans le cadre de CYCLONE, l'Institut Français de Bioinformatique (IFB) est en charge de la définition des besoins et cas d'utilisation en bioinformatique, la proposition de machines virtuelles prédéfinies (appliances) y répondant, la formation des utilisateurs à l'utilisation de ces appliances, la création de nouvelles appliances, et la proposition d'un portail à destination des utilisateurs.

Biosphère est un portail web à haut niveau d'abstraction qui permet à l'utilisateur de lancer les appliances "en un clic" sur différents clouds sans connaissance technique particulière, ou de configurer leur déploiement. Ce portail lui permet aussi de surveiller et gérer les déploiements,et la consommation des ressources. Le portail Biosphère contient un catalogue d'appliances bioinformatiques (RAINBio[1]), et permet d'explorer ces appliances et outils d'après les termes de l'ontologie EDAM et une recherche textuelle.

Afin d'assurer un authentification simple, de confiance, et utilisable autant dans le portail que dans les appliances, nous utilisons le gestionnaire d'identités développé dans le cadre du projet CYCLONE. Ce gestionnaire s'appuie sur la fédération d'identités académique européenne eduGAIN.

Une fois l'appliance choisie, il est possible de la lancer en un clic, et de configurer son déploiement. Lorsque l'appliance est basée sur une unique machine virtuelle (VM), il est possible de choisir la mémoire et le nombre de CPU; et dans le cas d'un cluster, le nombre de noeuds de calcul. Le portail Biosphère permet aussi de choisir le cloud utilisé, le cas échéant de lancer l'appliance sur plusieurs clouds, et de modifier ensuite dynamiquement le nombre de noeuds de calcul.

Dans les déploiements, l'authentification s'appuie sur la fédération d'identités eduGain, tant pour les connexions en terminal distant, qu'en bureau distant. L'isolation réseau d'un déploiement, par exemple pourun cluster de calcul, est assurée par le composant CNSMO, développé dans le cadre de CYCLONE, fournissant un réseau privé virtuel (VPN), basé sur le logiciel communautaire OpenVPN (exécuté dans un conteneur Docker).

Le portail Biosphère permet une gestion simplifiée des ressources utilisées. Afin de rendre le cloud accessible à tous, le portail Biosphère s'appuie sur les développements du projet CYCLONE. Il propose ainsi un haut niveau de sécurité réalisé par une authentification reposant sur la fédération d'identités académique européenne eduGAIN, et un placement des appliances au sein de réseaux isolés et sécurisés. Le portail permet aussi une allocation dynamique des ressources de calcul, et des déploiements multi-cloud. L'ensemble de ces fonctionalités étant proposées dans une interface se voulant intuitive.

References

[1] GROSJEAN Marie, HÉRIVEAU Claudia, JULLIEN Renaud, COLLIN Olivier, GIBRAT Jean-François, and BLANCHET Christophe. A RAINBio over the Life Sciences Cloud. Post-118 JOBIM 2015

Biodjango, an open framework for bioinformatics publishing

Ennys GHEYOUCHE, Stéphane TÉLETCHÉA

[1] UFIP, UMR 6286 CNRS Université de Nantes, 2 rue de la Houssinière, 44322 Nantes cedex 3

Corresponding Author: stephane.teletchea@univ-nantes.fr

Most of recent bioinformatics methods are available to the scientific community through web-based portals. Without technical knowledge, users can rapidly evaluate or use the method presented. This simplicity is often a tradeoff between the focused service provided and the integration with external resources. These web servers also lack of consistency for the presentation of services, for the organization of data and results.

With the advent of specialized editions of journals, some guidelines start to be set up for web services. Users nowadays expect to see a description of the method, a contact page, a simple form for job submission and a results page with sample examples. These web services are still mostly independent of external resources.

In this work, we present Biodjango, an integration of the Django framework for the presentation of web services for bioinformatics methods. Within this extensive framework, we provide integrated methods for linking external biological data to the analysis performed. On top of biopython or biodjango-specific methods, it is possible to handle Uniprot entries for protein annotation, Gene Ontology vocabulary, NCBI data and Bibliography management. For each biodjango application, a reference template is provided with examples so users building their own web service can rapidly adapt these applications for their needs. The application is made for modularity so users can pick only sub-parts of biodjango as required. To simplify the management of job submission, progression and display of results, biodjango offers their management with a simplified scheduler-like mechanism. To ease user adoption of biodjango, an extensive documentation is provided for a rapid set up of a biodjango-derived web service, this documentation will be updated regularly from the remarks and demands of the community.

We expect the biodjango project to accelerate web deployment of web services for the scientific community, so bioinformaticians will be able to dedicate more time for the development of innovative methods.

New Generation Phylogeny.fr: Refactoring Phylogeny.fr for Innovative Phylogenetic Services

Damien Correia[1], Vincent Lefort[2], Olivia Doppelt-Azeroual[3], Fabien Mareuil[3], Sarah Cohen-Boulakia[4] and Olivier Gascuel[2,3]

[1] Institut Français de Bioinformatique, IFB-Core - UMS 3601, Gif-sur-Yvette, France

[2] Laboratoire d'Informatique, de Robotique et de Microélectronique de Montpellier, LIRMM - UMR 5506 CNRS et Université de Montpellier, France

[3] Unité de Bioinformatique Evolutive et Hub Bioinformatique et Biostatistique, C3BI - USR 3756 Institut Pasteur et CNRS, Paris, France

[4] Laboratoire de Recherche en Informatique, LRI - CNRS UMR 8623 et Université Paris-Saclay, Orsay, France

Corresponding Author: damien.correia@pasteur.fr

1 Context

Phylogenetic analyses aim at reconstructing the evolutionary history of biological objects from molecules to species, and populations. Phylogeny plays a major role in highly diverse domains such as predicting biological functions or measuring the biodiversity of ecosystems. As a consequence, a plethora of approaches have been designed in various communities, resulting in a large number of programs available. Faced with the increasing need to perform phylogenetic analysis, and the difficulty for scientists to determine which program to use at each step of an analysis and how to combine the use of programs together, we designed Phylogeny.fr [1] almost ten years ago.

While Phylogeny.fr runs 50,000 data analysis per month, we are now faced with two major difficulties: (i) the series of programs used need to be updated and (ii) phylogeny.fr is used in contexts where its technology is reaching its limits (e.g., when simultaneously used by hundreds of students or when the server is used through batch scripts). Refactoring Phylogeny.fr is thus of paramount importance.

In the meantime, new solutions have emerged to help users manage their analyses: scientific workflow systems such as Galaxy [2] have reached a level of maturity making them particularly suitable for complex and large-scale analyses.

2 Sketch of the demonstration

In this demonstration, we introduce the first release of "NGPhylogeny.fr" (NGPhylogeny.pasteur.fr), developed under Affero GPL v2 licence within a python Web framework (Django), in which we have refactored phylogeny.fr by designing a scalable environment, an easy-to-use web interface and a series of modular Galaxy workflows to perform a large variety of phylogenetic analyses. All programs have been updated or replaced while some others have been added (such as *Noisy* to trim the alignment sites). Default parameter settings have also been revised. Our demonstration will be based on real datasets. We will show (i) how "NGPhylogeny.fr" can be used in a functional genomics context to quickly analyze large sets of protein superfamilies, (ii) how in depth studies can be launched and (iii) how "NGPhylogeny.fr" can be installed on a wide variety of configurations.

Acknowledgements: This project is supported by ANR-11-INBS-0013 - IFB.

References

[1] Phylogeny.fr: robust phylogenetic analysis for the non-specialist. Dereeper A, Guignon V, Blanc G, Audic S, Buffet S, Chevenet F, Dufayard JF, Guindon S, Lefort V, Lescot M, Claverie JM, Gascuel O. *Nucleic Acids Res.* 2008, 36(Web Server issue):W465-9

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

Conférence invitée

Franca Fraternali

Randall Division of Cellular and Molecular Biophysics, King's college, London, United Kingdom

Unraveling the Good and the Bad in Protein Networks: Functional versus Dysfunctional Interactions

In the last years, protein interactome comparisons have highlighted conserved modules that might represent common functional cores of ancestral origin. However, recent analyses of protein-protein interaction networks (PPINs) have led to a debate about the influence of the experimental method on the quality and biological relevance of these interaction data. It is crucial to know to what extent discrepancies between the networks of different species reflect sampling biases of the respective experimental methods, as opposed to topological features due to biological functionality. This requires new, precise and practical mathematical tools to quantify and compare the topological structures of networks at high resolution. To this end, we have studied the relationship between structured random graph ensembles and real biological signaling networks, focusing on the number of short loops in networks, which represent complexes in PPINs. By combination of a method for graph dynamics and an algorithm for loop counting, we estimated the relative importance of loops in biological networks compared to random graphs. We found that loops are a predominant feature of PPINs, suggesting that enrichment of their occurrence has a key functional role.

Nevertheless, one must keep in mind that not all the interactions between proteins result in a functional role that benefits the cell. One example is protein aggregation, resulting in neurotoxic assemblies that lead ultimately to cell death. We investigate in detail the case of interactions between fragments of the Prion protein (PrP) constituted by only the helices H2 and H3 of the entire protein. We have investigated the molecular mechanisms of the self-assembly process in solution by Molecular Dynamics. Our simulations show that this process occurs by assembly of small modules of four monomers that precede the creation of a base of six to eight H2H3 monomers; starting from this base, other H2H3 units attach to it in various configurations, assembling short filaments.

1) Chung SS, Pandini A, Annibale A, Coolen AC, Thomas NS, Fraternali F. Bridging topological and functional information in protein interaction networks by short loops profiling. Sci Rep. 2015; 5:8540.
2) Chakroun N, Fornili A, Prigent S, Kleinjung J, Dreiss CA, Rezaei H, Fraternali F. Decrypting Prion Protein Conversion into a β-Rich Conformer by Molecular Dynamics. J Chem Theory Comput. 2013; 5:2455-2465.
3) Chakroun N, Prigent S, Dreiss CA, Noinville S, Chapuis C, Fraternali F, Rezaei H. The oligomerization properties of prion protein are restricted to the H2H3 domain. FASEB J. 2010; 9:3222-31.

Use of cross-docking simulations for identification of protein-protein interactions sites: the case of proteins with multiple binding sites

Nathalie LAGARDE[1], Lydie VAMPARYS[1], Benoist LAURENT[1], Alessandra CARBONE[2,3] and Sophie SACQUIN-MORA[1]

[1] Laboratoire de Biochimie Théorique, CNRS UPR9080, Institut de Biologie Physico-Chimique, 13 rue Pierre et Marie Curie, 75005, Paris, France

[2] Laboratoire de Biologie Computationnelle et Quantitative, CNRS UMR7238, UPMC Univ-Paris 6, Sorbonne Université, 15 rue de l'Ecole de Médecine, 75006, Paris, France

[3] Institut Universitaire de France, 75005, Paris, France

Corresponding Author: lagarde@ibpc.fr, sacquin@ibpc.fr

1 Introduction

Understanding protein-protein interactions (PPI) is essential to decipher the mechanism of numerous biological functions. Some *in silico* methods can be used to investigate PPI. In particular, cross-docking simulations of large datasets of proteins can be used to predict interface residues [1-3]. In this study, we discussed the ability of the cross-docking method to detect the multiple binding sites on protein surfaces.

2 Methods

358 proteins extracted from 138 unique PDB structures were used for this study. The MAXDo algorithm [1] was used with a rigid-body docking approach and a reduced protein representation, a coarse-grain protein model developed by Zacharias [4]. Binding site predictions resulting from evolutionary sequence analysis produced with JET [5] were used to restrict the initial search space. For each surface residue, its Protein Interface Propensity (PIP) was computed and used to predict binding sites on the protein surface.

3 Results

For a large number of proteins, alternative interfaces different from the reference experimental interfaces were predicted. However, about 70 % of these interfaces were not false positives but correspond to interfaces with other partners (other chain of the same PDB not included in the database, nucleic acid molecule or homo-/hetero-dimerization interface). We compared the use of two different scoring schemes accounting for multiple binding sites, for evaluating the binding sites prediction. The first score was obtained by comparing the predicted interface with one single global reference experimental interface generated by concatening all the existing experimental interfaces. In the second score, the predicted interface was compared to each experimental interface separately, and only the interface associated with the best performance was kept.

4 Conclusion

Using cross-docking simulations on a large dataset of proteins, accurate binding sites preditions could be realized, including proteins which present multiple binding sites.

References

[1] Sophie Sacquin-Mora, Alessandra Carbone, and Richard Lavery. Identification of Protein Interaction Partners and Protein-Protein Interaction Sites. *Journal of Molecular Biology*, (382):1276-1289, 2008.

[2] Anne Lopes, Sophie Sacquin-Mora, Viktoriya Dimitrova, Elodie Laine, Yann Ponty and Alessandra Carbone. Protein-Protein Interactions in Crowded Environment: An Analysis via Cross-Docking Simulations and Evolutionary Information. *PLoS Computational Biology*, (9/12):e1003369,1-18, 2013.

[3] Lydie Vamparys, Benoist Laurent, Alessandra Carbone and Sophie Sacquin-Mora. Great interactions: How binding incorrect partners can teach us about protein recognition and function. *Proteins*, (84/10):1408-1421, 2016.

[4] Martin Zacharias. Protein-protein docking with a reduced protein model accounting for side-chain flexibility. *Protein Science*, (12):1271-1282, 2003.

[5] Stefan Engelen, Ladislas A Trojan, Sophie Sacquin-Mora, Richard Lavery and Alessandra Carbone. Joint Evolutionary Trees: A Large-Scale Method To Predict Protein Interfaces Based on Sequence Sampling. *PLoS Computational Biology*, (5/1):1-17, 2009.

In silico developments for the study of glycosylation applied to extracellular matrix proteins

Camille BESANÇON[1], Alexandre GUILLOT[1], Sébastien BLAISE[1], Manuel DAUCHEZ[1,2], Jessica JONQUET[1], Nicolas BELLOY[1,2], , Stéphanie BAUD[1,2]

[1] UMR CNRS/URCA N° 7369, Université de Reims Champagne-Ardenne, UFR Sciences Exactes et Naturelles, Moulin de la Housse, 51687 Reims Cedex 2,France

[2] Plateau de Modélisation Moléculaire Multiéchelle, Université de Reims Champagne-Ardenne, UFR Sciences Exactes et Naturelles, Moulin de la Housse, 51687 Reims Cedex 2,France

Corresponding Author: camille.besancon@etudiant.univ-reims.fr

N-glycosylations play an important role in protein functions and some alterations of glycosylation such as sialic-acid hydrolysis can alter protein activity. Some alterations are also known as pathological markers. Considering their biological importance, it is essential to study glycosylated proteins. However, *in vitro* study of N-glycans can be very difficult because of the structural diversity and the many reactive groups of the glycan chains. Molecular dynamics can be a useful tool to overcome this problem and give access to conformational informations through exhaustive sampling. At the beginning of this project, we lacked a tool, enabling to realize in a single workflow, from the building of a glycan structure to the analysis of the trajectories. To adress this issue, we want to use these studies to propose a complete tool for the building, the simulation and the analysis of glycosylated proteins that will be implemented in GROMACS.

The present work is based on a previous study from our laboratory: we investigated the structural influence of sialic-acid on N-glycan combining *in vitro* and *in silico* approaches. A new method to evaluate the conformational landscape has been developed by projecting 3D vectors on a geometric plan. We aligned the glycan core on the Z axis in an orthogonal space before projecting the position of the glycan branches on the XY plan. With this method, along with the clustering of representative structures and the measurement of dihedral angles values of the glycosidic linkage, it was possible to show the importance of sialic acids and their role in the structural flexibility of N-glycans [1].

To investigate the structural influence of glycosylation on protein structure, we focus on the glycosylated Insulin Receptor (IR). It was previously shown that the hydrolysis of sialic acids on the IR by neuraminidase-1 sialidase induced insulin resistance and disrupted cell glucose uptake [2]. Simultaneously, we will work on new methods to specifically analyze protein / glycan complexes. We're aiming at developing a method for the measurement of the protein surface covered by the glycan in order to investigate the structural role of the sugar chains on protein surfaces.

At this time, using the online tool CHARMM-GUI [3] to generate the structures, we have created a first glycan library that will be implemented in our GROMACS module. We first concentrated on biologically relevant N-glycans (pathological markers identified by literature search [4,5]). We then started molecular dynamics simulations on some of these glycans. Our aim is to further our analysis and our results on the matter of N-glycan structure and flexibility. With this data set, we will be able to start implementing the GROMACS module and test our analysis methods. From this work, we aim at providing a complete tool paired with specific analysis methods thus enabling the building, the simulation of glycosylated proteins.

References

[1] Guillot, A., Dauchez, M., Belloy, N., Jonquet, J., Duca, L., Romier, B., Maurice, P., Debelle, L., Martiny, L., Durlach, V., et al. Impact of sialic acids on the molecular dynamic of bi-antennary and tri-antennary glycans. *Sci. Rep.* 6, 35666. ,2016.

[2] Hayes, G. R. & Lockwood, D. H. The role of cell surface sialic acid in insulin receptor function and insulin action. *J. Biol. Chem.*261, 2791–2798 (1986).

[3] Jo, S., Song, K.C., Desaire, H., MacKerell, A.D., and Im, W. (2011). Glycan Reader: Automated Sugar Identification and Simulation Preparation for Carbohydrates and Glycoproteins. *J. Comput. Chem. 32*, 3135–3141.

[4] Kawar, Z.S., Haslam, S.M., Morris, H.R., Dell, A., and Cummings, R.D. (2005). Novel poly-GalNAcbeta1-4GlcNAc (LacdiNAc) and fucosylated poly-LacdiNAc N-glycans from mammalian cells expressing beta1,4 Nacetylgalactosaminyltransferase and alpha1,3-fucosyltransferase. *J. Biol. Chem*. *280*, 12810–12819.

[5] Lee, L.Y., Thaysen-Andersen, M., Baker, M.S., Packer, N.H., Hancock, W.S., and Fanayan, S. (2014). Comprehensive Nglycome profiling of cultured human epithelial breast cells identifies unique secretome N-glycosylation signatures enabling tumorigenic subtype classification. *J. Proteome Res. 13*, 4783–4795.

HG-CoLoR: Hybrid Graph for the error Correction of Long Reads

Pierre MORISSE, Thierry LECROQ and Arnaud LEFEBVRE
Normandie-Université, UNIROUEN, LITIS EA4108, 76821, Mt-St-Aignan, France

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

Abstract *The recent rise of long read sequencing technologies allows the solving of assembly problems for large and complex genomes that were, until then, unsolvable with the use of short read sequencing technologies alone. Despite the fact that they can reach lengths of tens of kbps, these long reads are very noisy, and can reach an error rate as high as 30%, involving mandatory error correction before using them to efficiently solve assembly problems. However, as the vast majority of these errors are insertions and deletions, classical error correction tools developed for short reads, which mainly focus on substitution errors, are not effective for correcting long reads. Therefore, several new methods specifically designed for long read error correction have recently been developed. In particular, NaS, instead of directly correcting the long reads, proposes to use them as templates in order to produce assemblies of related accurate short reads, and use them as corrections. Following this idea, we introduce HG-CoLoR, a new tool for the production of such corrected long reads, that gets rid of the need to align the short reads against each other, which is the bottleneck from NaS. Indeed, HG-CoLoR focuses on a seed-and-extend approach based on a hybrid graph built from the short reads. Our experiments show that, while producing comparable results both in terms of length and accuracy of the corrected long reads, HG-CoLoR is several times faster than NaS, and also yields better assembly results than other state-of-the-art long read hybrid error correction methods. HG-CoLoR is available from https://github.com/pierre-morisse/HG-CoLoR.*

Keywords NGS, long reads, correction, assembly

1 Introduction

Since a few years, long read sequencing technologies are being developed, and allow the solving of assembly problems for large and complex genomes that were, until then, unsolvable with the use of short reads sequencing technologies alone. The two major actors of these long read sequencing technologies are Pacific Biosciences and Oxford Nanopore, which, with the release of the MinION device, allows a low-cost and easy long read sequencing.

However, even though long reads can reach lengths of tens of kbps, they also reach a very high error rate of around 15% for Pacific Biosciences, and up to 30% for Oxford Nanopore, the vast majority of these errors being insertions and deletions. Due to this high error rate, correcting these long reads before using them to efficiently solve assembly problems is mandatory. Many methods are available for short read correction, but these methods are not applicable to long reads, on the one hand because of their much higher error rate, and on the other hand, because most of the error correction tools for short reads focus on substitution errors, the dominant error type in Illumina data, whereas insertions and deletions are more common in long reads.

Recently, several methods for long read correction have been developed. These methods can be divided into two main categories: either the long reads are selfcorrected by aligning them against each other (HGAP [1], PBcR [2]), or either a hybrid strategy is adopted, in which the long reads are corrected with the help of accurate short reads (LSC [3], proovread [4], CoLoRMap [5]). de Bruijn graph [6] based methods, where the long reads are mapped on the graph, and erroneous regions corrected by traversing its paths, also started to develop recently, in the hybrid case (LoRDEC [7], Jabba [8]), as well as in the non-hybrid case (LoRMA [9]).

NaS [10], instead of locally correcting the long reads, uses them as templates to produce corrected long reads from assemblies of related accurate short reads. The short reads are mapped both on these templates, and against each other, in order to associate a subset of related short reads to each template. Each subset in then assembled, and the obtained contig is used as the correction of the associated template.

In this paper, we introduce HG-CoLoR, a new long read hybrid error correction method that combines both the main idea from NaS to produce corrected long reads from assemblies of accurate short reads, and the use

of a graph, in order to get rid of the time consuming step of aligning all the short reads against each other. HG-CoLoR indeed focuses on a seed-and-extend approach where the seeds, which are short reads that align correctly on the long reads, are used as anchor points on a graph that is traversed in order to link them together and to produce the corrected long reads. This graph is actually a hybrid structure between a de Bruijn graph and an overlap graph [12]. It is defined from the short reads' k-mers, and allows to compute perfect overlaps of variable length between these k-mers. It is not explicitly built, but the use of PgSA [11] allows to simulate its traversal.

Our experiments show that, while producing comparable results both in terms of length and accuracy of the corrected long reads, HG-CoLoR is several times faster than NaS, and also yields better assembly results than other state-of-the-art long read hybrid error correction methods.

For the sake of understanding, we first give an overview of NaS, and describe our hybrid graph and the way PgSA allows its traversal, before introducing HG-CoLoR.

2 NaS Overview

NaS is a hybrid method for the error correction of long reads that, unlike other methods, uses long reads as templates rather than directly correcting them. Short reads are mapped both on these templates and against each other, in order to gather different subsets of short reads, each subset related to one given template. Each subset is then assembled, and the produced contig is used as the correction of the related template. More precisely, a corrected long read is produced from a template as follows.

First, the short reads are aligned on the template using BLAT [13] in fast mode, or LAST [14] in sensitive mode, in order to find seeds, which are short reads that align correctly on the template. Then, once these seeds have been found, all the short reads are aligned against each other, and similar reads, which are reads that share a certain number of non-overlapping k-mers with the seeds, are recruited with the help of Commet [15]. Finally, the obtained subset of short reads is assembled using Newbler (unpublished), and a contig is produced, and used as the correction of the initial template

The reads recruitment is the most crucial step of the method, as it allows to retrieve short reads corresponding to low quality regions of the template. However, this step is also the bottleneck of the whole NaS pipeline, as it is responsible for 70% of the total runtime on average.

NaS is able to generate corrected long reads up to 60 kbps, that align entirely on the reference genome and that span repetitive regions. On average, the accuracy of the corrected long reads produced by NaS reaches 99.75%, without any significant length drop compared to the input long reads. Moreover, these corrected long reads also yield highly contiguous assembly results, and thus show that focusing on the production of corrected long reads from assemblies of accurate short reads, rather than on the local correction of the long reads, is an interesting alternative to classical long read hybrid error correction.

3 Hybrid graph

As previously mentioned, the graph used by HG-CoLoR is a hybrid structure between a de Bruijn graph and an overlap graph. This hybrid graph is not explicitly built, but can be traversed with the help of PgSA, which is a data structure that allows the indexing of a set of reads of constant length, in order to answer different queries, for a given string f. For place sake, we do not detail how the index is built, the complete list of queries, nor how they are processed. For more details, one can refer to [11]. We simply mention that PgSA supports querying for variable lengths of f without recomputing the index, and that one of the queries returns the positions of all the occurrences of f in the different reads of the set.

This way, using PgSA to index a set of reads, and looping over the aforementioned query, allows to compute perfect overlaps of variable length between the reads, thus simulating the use of an overlap graph. In the same fashion, indexing the k-mers from a set of reads, and looping over the aforementioned query, fixing the length of the queries strings as $k-1$, allows to compute perfect overlaps of length $k-1$ between the k-mers, thus simulating the use of a de Bruijn graph. However, indexing the k-mers from a set of reads, and looping over the aforementioned query, of course, also allows to compute perfect overlaps of variable length between the different k-mers, thus simulating the use of a hybrid structure between a de Bruijn graph and an overlap graph. To the best of our knowledge, this is the first time such a structure is mentioned. For better understanding, an example of a simple graph is given in Figure 1.

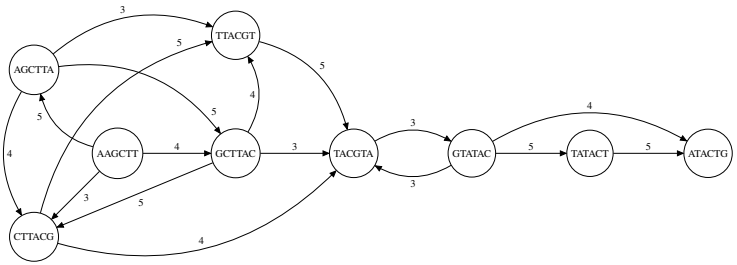


Fig. 1. A simple example of our graph, when fixing the length of the k-mers to 6, computing overlaps of minimum length 3, and building from the three following reads: AAGCTTAC, CTTACGTA, GTATACTG. Numbers on the edges of the graph represent the overlap length between the k-mers.

4 HG-CoLoR description

HG-CoLoR, like NaS, aims to use erroneous long reads as templates, and to produce corrected long reads from assemblies of short reads related to these templates. However, its main objective is to get rid of the time consuming step of reads recruiting, that requires the mapping of all the short reads against each other. To do so, it focuses on a seed-and-extend approach where the seeds are found in the same way as NaS, and where the k-mers from the short reads, and their reverse-complements, are indexed with PgSA, to allow the traversal of the previously described graph. This graph is traversed, in order to extend and link together the seeds, used as anchor points, by directly assembling the short reads' k-mers during the traversal. HG-CoLoR's workflow is summarized in Figure 2, and its four main steps are described below.

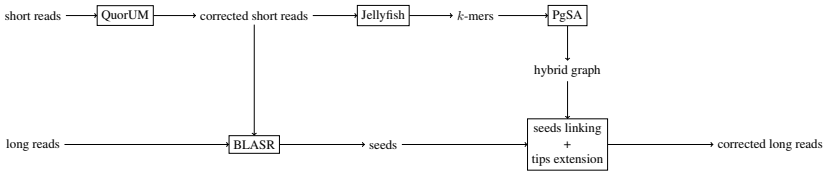


Fig. 2. HG-CoLoR's workflow. First, the short reads are corrected in order to get rid of as much sequencing errors as possible. Then, all the k-mers from the corrected short reads, and their reverse-complements, are obtained with Jellyfish, and indexed with PgSA, to allow the traversal of the graph. The corrected short reads are aligned on the long reads with BLASR to find seeds, and each long read is then considered as a template, and processed independently. For a given template, the graph is traversed in order to extend and link together the associated seeds, used as anchor points. Then, the tips of the sequence obtained after the seeds linking step are extended in both directions by traversing the graph, to reach the initial template's borders. Finally, the corrected long read is output.

4.1 Short reads correction and indexing

Even though short reads are very accurate prior to any correction, as HG-CoLoR seeks to arrange their k-mers into a graph structure, and traverse it to extend and link the seeds together, it needs to get rid of as much sequencing errors as it can in this data. Thus, prior to any other step, the short reads are corrected with the help of QuorUM [16], which is able to provide a good raise of the accuracy in very little time. Then, the k-mers from the corrected short reads, and their reverse-complements, are extracted with Jellyfish [17], and indexed with PgSA, in order to allow the traversal of the graph during the following steps.

4.2 Seeds retrieving and merging

Like with NaS, the seeds are found by mapping the corrected short reads on the long reads, used as templates. This is done with the help of BLASR [18], an alignment tool specifically designed to align long reads dominated by insertion and deletion errors. Then, each template is processed independently, and two phases of analyze and merging are applied to the associated seeds. First, if the mapping positions of a given couple of

seeds imply that they overlap on the template over a sufficient length, their assumed overlapping sequences are compared, and the two seeds are merged accordingly. If the mapping positions indicate that the two seeds do overlap on the template, but not over a sufficient length, or if the assumed overlapping sequences do not coincide, only the seed with the best alignment score is kept. Then, once all the seeds with overlapping mapping positions have been merged or filtered out, sequence overlaps between consecutive seeds are computed. As in the previous step, if a given seed overlaps the following one over a sufficient length, the two seeds are merged.

4.3 Seeds linking

Once the seeds have been found and merged for all of the templates, HG-CoLoR once again processes each template independently and attempts to link the related seeds together by considering them as couples, and traversing the graph. The rightmost k-mer of the left seed (source) and the leftmost k-mer of the right seed (destination) are used as anchor points, and the source is extended with perfectly overlapping k-mers from the corrected short reads, found by following the paths of the graph, until the destination is reached. When facing branching paths, every possible path is explored with the use of backtracking, to find the one that will allow correct linking of the source to destination. Of course, HG-CoLoR explores these different paths in decreasing order of the overlaps lengths, which means that edges representing longer overlaps are always explored before those representing shorter ones. It also only explores edges that represent overlaps that are longer than a defined minimum length. Moreover, as short reads from a different region of the reference genome can align on the template and can be used as seeds, thus leading to impossible linkings, a threshold on the maximum number of backtracks is set, to avoid useless important runtime and intensive computation.

If this threshold is reached, and no path has been found to link the source to the destination, the current linking iteration is given up. When such a situation occurs, two different cases have to be taken into account. In the first case, if no seeds have been linked so far, the current source is simply ignored, and a new linking iteration is computed for the next couple of seeds. In the second case, if seeds have already been linked previously, the source remains the same, the destination seed that could not be reached is ignored, and the destination is defined as the next seed for the next linking iteration. An illustration of these two different cases is given in Figure 3.

However, in the second case, as this process of skipping a seed in the middle of the template can provoke an important number of failed linking attempts, if seeds from a wrong region are present in great proportion on the template, a threshold on the maximum number of seeds that can be skipped is set. Once this threshold is reached, if the sequence obtained from the previously linked seeds could not be extended to reach one of the remaining seeds, HG-CoLoR attempts to produce a fragmented corrected long read: the part corresponding to the seeds linked so far is output, and the graph is traversed again, in order to try to link the remaining seeds together, independently of the previous part.

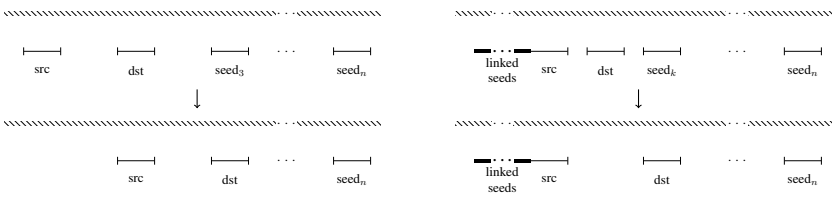


Fig. 3. Illustration of the different cases of seed skips. Hatched lines represent the templates, standard segments represent the seeds, and bold segments represent the sequences obtained from the previously linked seeds. First case (left): No seeds have been linked so far, the current source seed is simply ignored, and both the source and the destination are moved to the next couple of seeds. Second case (right): Seeds have already been linked previously, the source remains the same, the destination seed that could not be reached is ignored, and the destination is defined as the next seed.

4.4 Tips extension

Finally, it is obvious that the seeds do not always map right at the beginning and until the end of the templates. Thus, in order to get as close as possible to the original templates' lengths, once all the seeds of a given template have been linked, HG-CoLoR keeps on traversing the graph and extending the tips of the produced corrected long read, on the left of the leftmost seed, and on the right of the rightmost seed, until they reach the template's borders, or a branching path. Indeed, in the case of tips extension, when facing a

branching path, HG-CoLoR has no clue as to which path to chose and continue the extension with, nor any anchor points, unlike when it attempts to link two seeds together. Therefore, backtracking is useless and the extension is simply stopped when such a situation occurs. In the case of fragmented corrected long reads, as HG-CoLoR can not properly rely on the template's borders, every fragment is extended until a branching path is reached.

5 Results and discussion

We compare the quality of our corrected long reads with those produced by NaS, and also with those produced by two other state-of-the-art hybrid error correction methods, namely CoLoRMap and Jabba (which is more recent than LoRDEC). We compare the results both in terms of alignment identity of the corrected long reads, and in terms of quality of the assemblies that could be generated from these reads.

5.1 Parameters

We ran multiple rounds of correction with HG-CoLoR on the different datasets to experiment with the parameters, and find the combination that would produce the best results. Thereby, we found that a k-mer size of 64 for the graph construction yielded the best compromise between accuracy, genome coverage, and average length of the output corrected long reads. The minimum overlap length to allow the merging of two seeds during the second step was set to 63, accordingly to the k-mer size chosen for the graph construction. The minimum overlap length allowed to explore an edge during the graph traversal was set to 59, as decreasing it more yielded unsatisfying results, and increasing it would make our graph closer to an actual de Bruijn graph than to the hybrid graph it's supposed to be. The maximum number of backtracks was set to 1,125, as decreasing it more drastically impacted the quality of the produced corrected long reads, and increasing it, even to very large values, barely yielded better results, but greatly increased the runtime. For the same reason, the maximum number of seed skips was set to 5. For the mapping of the short reads on the long reads, BLASR was used with default parameters except for bestn, that was set to 30 instead of 10. Yet again, increasing this parameter to larger values only impacted the runtime, and did not improve the correction results enough to be interesting, while decreasing it induced a drop of the number of output corrected long reads. Finally, GNU Parallel [19] was used to allow HG-CoLoR to run on multiple processes. CoLoRMap was run with default parameters. Following the authors' recommendations, before running Jabba, the short reads were corrected with Karect [20], and the de Bruijn graph was constructed and corrected with Brownie (unpublished). As choosing the same value as the one used for HG-CoLoR led to worse results, a k-mer size of 75 was chosen for the graph construction, as recommended by the authors. All tools were run with 16 processes.

5.2 Datasets

As we mainly seek to compare our results with NaS, we use the same data to allow a better comparison. This data is composed of both long Oxford Nanopore reads and short Illumina reads for three different genomes: *Acinetobacter baylyi*, *Escherichia coli*, and *Saccharomyces cerevisae*. Details are given in Table 1.

Dataset	Reference genome				Oxford Nanopore data			Illumina data		
	Name	Strain	Reference sequence	Genome size	# Reads	Average length	Coverage	# Reads	Read length	Coverage
A. baylyi	*A. baylyi*	ADP1	CR543861	3.6 Mbp	89,011	4,284	44x	900,000	250	50x
E. coli	*E. coli*	K-12 substr. MG1655	NC_000913	4.6 Mbp	22,270	5,999	28x	775,500	300	50x
Yeast	*S. cerevisae*	W303	scf7180000000084-113	12.4 Mbp	205,923	5,698	31x	2,500,000	250	50x

Tab. 1. Description of the datasets used in our experiments. Both MinION and Illumina data are available from the Genoscope's website `http://www.genoscope.cns.fr/externe/nas/datasets.html`.

5.3 Alignment-based comparison

The previously described long reads datasets were aligned with Last prior to any correction. The four different correction tools were then applied, and the obtained corrected long reads were aligned with BWA mem [21]. Results are given in Table 2 and discussed below.

We notice that, unlike the other methods, CoLoRMap output all the long reads and not only the ones it managed to correct. As the reads that could be corrected were not tagged in any way and could therefore not be extracted, it appears that CoLoRMap performed the worst correction, and did not manage to improve the accuracy of the long reads at all, except for the *E. coli* dataset. These poor results are probably due to the

Dataset	Method	# Reads	Average length	Cumulatize size	# Aligned reads	Average identity	# Error-free reads	Genome coverage	Runtime
A. baylyi	Original	89,011	4,284	381,365,755	29,954	70.09%	0	100%	N/A
	CoLoRMap	89,011	4,355	387,609,994	18,085	67.93%	2	100%	14h33min
	Jabba	17,476	10,260	179,309,738	17,476	99.40%	16,893	99.80%	12min30
	NaS (fast)	24,063	8,840	212,707,189	24,063	99.82%	22,984	100%	94h18min
	NaS (sensitive)	28,492	9,530	271,526,778	28,492	99.83%	27,190	100%	128h55min
	HG-CoLoR	23,465	11,137	261,327,970	23,461	99.44%	20,906	100 %	21h08min
E. coli	Original	22,270	5,999	133,607,392	22,170	79.46%	0	100%	N/A
	CoLoRMap	22,270	6,219	138,489,144	21,784	89.02%	152	100%	8h26min
	Jabba	22,065	5,794	127,848,525	22,065	99.81%	21,850	99.41%	12min56
	NaS (fast)	21,818	7,926	172,918,739	21,818	99.86%	20,383	100%	72h02min
	NaS (sensitive)	22,144	8,307	183,958,832	22,144	99.86%	20,627	100%	81h30min
	HG-CoLoR	22,549	5,897	132,979,813	22,549	99.59%	19,676	100%	15h15min
Yeast	Original	205,923	5,698	1,173,389,509	68,215	55.49%	0	99.90%	N/A
	CoLoRMap	205,923	5,737	1,181,298,941	40,530	39.93%	23	99.40%	37h36min
	Jabba	36,958	6,613	244,402,749	36,855	99.55%	34,028	93.21%	44min05
	NaS (fast)	71,793	5,938	426,326,355	71,664	99.59%	59,788	98.70%	-
	NaS (sensitive)	85,432	6,770	578,351,588	85,288	99.53%	69,816	99.17%	-
	HG-CoLoR	71,518	6,604	472,306,800	71,393	99.17%	55,357	98.39%	99h16min

Tab. 2. Runtime and statistics of the long reads, before and after correction by the different tools. NaS runtimes are omitted for the Yeast dataset because the results did not compute in 16 days. NaS corrected reads for this dataset were obtained from the Genoscope website.

fact that only a few reads could be corrected, as CoLoRMap is designed to correct long reads from Pacific Biosciences, that have an error rate of about 15%, whereas the long reads used in our experiments were from Oxford Nanopore, and reached an error of at least 30% for the two other datasets.

Jabba clearly performed the best when it comes to runtime, outperforming all the other tools by several orders of magnitude. It also produced corrected long reads that aligned with a high identity, a great proportion of them aligning with no error. However, although highly accurate, these corrected long reads did not manage to completely cover any of the studied reference genomes.

When it comes to this point, only NaS and HG-CoLoR managed to cover the whole reference genomes with high identity, except for Yeast, due to the fact that even the original long reads did not cover the whole genome. Moreover, HG-CoLoR outperforming Jabba in terms of genome coverage also tends to underline the usefulness of our hybrid graph, showing that it seems to resolve the different regions of the reference genomes better than a classical de Bruijn graph, when the short reads coverage is locally insufficient.

On the three datasets, NaS yielded more corrected long reads than HG-CoLoR, both in fast and sensitive mode. The slight advantage of HG-CoLoR on the *E. coli* dataset comes from the production of fragmented corrected long reads, rather than from a greater number of processed templates. In both modes, the corrected long reads produced by NaS also aligned with a slightly higher identity than those produce by HG-CoLoR, and a greater proportion was therefore error-free. As for the average length and the cumulative size of the corrected long reads, HG-CoLoR performances were highly similar to NaS's, except on the *E.coli* dataset, where the advantage of NaS is probably due to the high quality of the original templates, and to the fact that it can recruit short reads outside of the templates, while HG-CoLoR stops once the borders are reached. However, despite its slight disadvantage on the aforementioned metrics, HG-CoLoR was at least four times faster than NaS, even in fast mode.

5.4 Assembly-based comparison

All the corrected long reads datasets previously described were assembled using Canu [22], without the correction and trimming steps. The following parameters were used for the assembly of all the datasets: OvlMerSize=17, MhapMerSize=17, OvlMerDistinct=0.9925, OvlMerTotal=0.9925. The correctedErrorRate parameter was tuned independently for each dataset. It was set to 0.07 for *A. baylyi*, to 0.085 for *E. coli* and to 0.125 for Yeast. Results are given in Table 3 and discussed below.

In agreement with what we observed in Table 2, the low accuracy of the long reads corrected by CoLoRMap resulted in impossible assemblies. Only the corrected long reads of the *E. coli* dataset could be assembled, due to their original high accuracy, but the generated assembly did not cover the whole genome, and displayed the worst identity among all the other assemblies.

As for Jabba, the fact that the corrected long reads did not manage to cover the whole reference genomes resulted in highly fragmented assemblies, that could not resolve large regions of the reference genomes. As a

Dataset	Method	# Reads	Coverage	# Expected contigs	# Obtained contigs	Genome coverage	Identity
A. baylyi	CoLoRMap	89,011	44x	1	-	-	-
	Jabba	17,476	50x	1	13	89.43%	99.93%
	NaS (fast)	24,063	59x	1	1	100 %	99.99 %
	NaS (sensitive)	28,492	75x	1	2	99.72%	99.98%
	HG-CoLoR	23,465	73x	1	1	99.97%	99.93%
E. coli	CoLoRMap	22,270	28x	1	29	97,74%	99.81%
	Jabba	22,065	28x	1	41	95.76%	99.92%
	NaS (fast)	21,818	37x	1	1	99.90 %	99.99%
	NaS (sensitive)	22,144	40x	1	2	100%	99.99%
	HG-CoLoR	22,549	29x	1	2	99.95%	99.95%
Yeast	CoLoRMap	205,923	14x	30	-	-	-
	Jabba	36,958	21x	30	134	70.52%	99.83%
	NaS (fast)	71,793	35x	30	123	97.44%	99.77%
	NaS (sensitive)	85,432	47x	30	123	96.98%	99.80%
	HG-CoLoR	71,518	39x	30	108	92.19%	99.61%

Tab. 3. Statistics of the assemblies that were generated from the long reads, after correction by the different tools. CoLoRMap results are ommitted for the *A. baylyi* and Yeast datasets, because Canu did not manage to assemble the sets of corrected long reads.

result, long reads corrected by Jabba yielded the least covering assemblies, despite their high average length and high accuracy. This underlines the fact that, although it is extremely fast, Jabba does not seem to be adapted for correcting long reads prior to an assembly.

Surprisingly, for all the datasets, the sensitive mode of NaS produced corrected long reads that resulted in slightly less satisfying assemblies than the fast mode. However, the difference was not significant, and adapting the parameters of Canu to match the corrected long reads produced in sensitive mode addressed this issue.

Therefore, only the corrected long reads produced by NaS and HG-CoLoR could be assembled into a decent number of contigs, covering the reference genomes well, and with a high identity. However, for the Yeast dataset, none of these two tools managed to produce corrected long reads allowing to get close to the expected number of contigs, nor to the full genome coverage. This is probably due to the fact that the original long reads were of really poor quality, displaying an error rate of almost 45%, and did not cover the whole genome. They were indeed sequenced with an old chemistry, and it is more than likely that, with long reads from a more recent one as templates, both NaS and HG-CoLoR could produce corrected long reads that would greatly reduce the number of contigs and increase the genome coverage of the assembly.

6 Conclusion

We described HG-CoLoR, a new hybrid method for the error correction of long reads, that, like NaS, uses long reads as templates and focuses on the production of corrected long reads from assemblies of accurate short reads, rather than on the local correction of the input long reads. Our method, instead of aligning the short reads against each other in a recruiting step, like NaS, focuses on a seed-and-extend approach and introduces a brand new idea of using a hybrid structure between a de Bruijn graph and an overlap graph. This graph, which is defined from the short reads' k-mers, and traversed with PgSA, is used to extend and link together the seeds, which are short reads that align correctly on the input long reads, by a simple traversal, using them as anchor points. Therefore, the corrected long reads are produced by directly assembling the short reads' k-mers during the traversal, without using any other proper assembly tool.

We tested this new method and compared it with NaS, CoLoRMap and Jabba on Oxford Nanopore long reads from three different genomes, namely *A. baylyi*, *E. coli*, and *S. cerevisae*. On these three datasets, HG-CoLoR yielded results that compared well with NaS, while being several times faster, CoLoRMap produced corrected long reads of poor quality, and Jabba, while being the fastest tool, produced accurate corrected long reads that however did not cover the whole reference genomes. As a result, only the corrected long reads produced by NaS and HG-CoLoR could be assembled into a decent number of contigs, covering well the reference genomes, although NaS outperformed HG-CoLoR on the *S. cerevisae* dataset.

The development of this method shows that, when having anchor points, the previously introduced hybrid

graph can prove useful for hybrid error correction of long reads, and can even yield better results than a classical de Bruijn graph. For future works, it could be interesting to focus more on this graph, and directly build it instead of simulating its traversal with PgSA, in order to directly map the long reads on the graph, like Jabba, thus skipping the alignment step of the short reads on the long reads, and reducing the runtime.

Acknowledgements

The authors would like to thank the Genoscope team for the availability of all the data used in this paper. This work has been partially supported by the Defi MASTODONS C3G project from CNRS.

References

[1] 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(6):563–569, 2013.

[2] Konstantin Berlin, Sergey Koren, Chen-Shan Chin, James P Drake, Jane M Landolin, and Adam M Phillippy. Assembling large genomes with single-molecule sequencing and locality-sensitive hashing. *Nature biotechnology*, 33(6):623–630, 2015.

[3] 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.

[4] 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.

[5] Ehsan Haghshenas, Faraz Hach, S Cenk Sahinalp, and Cedric Chauve. CoLoRMap: Correcting Long Reads by Mapping short reads. *Bioinformatics*, 32(17):i545–i551, 2016.

[6] Nicolaas Govert de Bruijn. A combinatorial problem. *Koninklijke Nederlandse Akademie van Wetenschappen te Amsterdam*, 49(7):758–764, 1946.

[7] Leena Salmela and Eric Rivals. LoRDEC: Accurate and efficient long read error correction. *Bioinformatics*, 30(24):3506–3514, 2014.

[8] Giles Miclotte, Mahdi Heydari, Piet Demeester, Stephane Rombauts, Yves Van de Peer, Pieter Audenaert, and Jan Fostier. Jabba: hybrid error correction for long sequencing reads. *Algorithms Mol Biol*, 11:10, 2016.

[9] Leena Salmela, Riku Walve, Eric Rivals, and Esko Ukkonen. Accurate selfcorrection of errors in long reads using de Bruijn graphs. *Bioinformatics*, 33(6):799–806, 2017.

[10] Mohammed-Amin Madoui, Stefan Engelen, Corinne Cruaud, Caroline Belser, Laurie Bertrand, Adriana Alberti, Arnaud Lemainque, Patrick Wincker, and Jean-Marc Aury. Genome assembly using Nanopore-guided long and error-free DNA reads. *BMC Genomics*, 16:327, 2015.

[11] Tomasz Kowalski, Szymon Grabowski, and Sebastian Deorowicz. Indexing arbitrary-length k-mers in sequencing reads. *PLoS ONE*, 10(7):1–14, 2015.

[12] Andrzej Ehrenfeucht, Tero Harju, Ion Petre, David M Prescott, and Grzegorz Rozenberg. *Overlap Graphs*, pages 99–108. Springer Berlin Heidelberg, Berlin, Heidelberg, 2004.

[13] W James Kent. BLAT — The BLAST -Like Alignment Tool. *Genome research*, 12:656–664, 2002.

[14] Szymon M Kielbasa, Raymond Wan, Kengo Sato, Szymon M Kiebasa, Paul Horton, and Martin C Frith. Adaptive seeds tame genomic sequence comparison. *Genome Research*, 21(3):487–493, 2011.

[15] Nicolas Maillet, Guillaume Collet, Thomas Vannier, Dominique Lavenier, and Pierre Peterlongo. Commet: Comparing and combining multiple metagenomic datasets. In *IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, 2014.

[16] Guillaume Marçais, James A Yorke, and Aleksey Zimin. QuorUM: An Error Corrector for Illumina Reads. *PLOS ONE*, 10(6):1–13, 2015.

[17] Guillaume Marçais and Carl Kingsford. A fast, lock-free approach for efficient parallel counting of occurrences of k-mers. *Bioinformatics*, 27(6):764–770, 2011.

[18] Mark J Chaisson and Glenn Tesler. Mapping single molecule sequencing reads using basic local alignment with successive refinement (BLASR): application and theory. *BMC bioinformatics*, 13(1):238, 2012.

[19] Ole Tange. GNU Parallel - The Command-Line Power Tool. *;login: The USENIX Magazine*, 36(1):42–47, 2011.

[20] Amin Allam, Panos Kalnis, and Victor Solovyev. Karect: accurate correction of substitution, insertion and deletion errors for next-generation sequencing data. *Bioinformatics*, 31(21):3421–3428, 2015.

[21] Heng Li and Richard Durbin. Fast and accurate long-read alignment with Burrows-Wheeler transform. *Bioinformatics*, 26(5):589–595, 2010.

[22] Sergey Koren, Brian P. Walenz, Konstantin Berlin, Jason R. Miller, and Adam M. Phillippy. Canu: scalable and accurate long-read assembly via adaptive k-mer weighting and repeat separation. *bioRxiv*, 2016.

PhylOligo: a package to identify contaminant or untargeted organism sequences in genome assemblies.

Ludovic Mallet [1†], Tristan Bitard-Feildel [2†], Franck Cerutti [1] and Hélène Chiapello [1]
[1] INRA UR875, Unité Mathématiques et Informatique Appliquées de Toulouse (MIAT), Auzeville, 31326 Castanet-Tolosan, France
[2] CNRS UMR7590, Sorbonne Universités, Université Pierre et Marie Curie - Paris 6 - MNHN - IRD - IUC, Paris, France.

Corresponding author: ludovic.mallet@inra.fr

Reference paper: **Mallet *et al.* (2017) PhylOligo: a package to identify contaminant or untargeted organism sequences in genome assemblies. *in rev*.**

Abstract *Genome sequencing projects sometimes uncover more organisms than expected, especially for complex and/or non-model organisms. It is therefore useful to develop software to identify mix of organisms from genome sequence assemblies. We developed a suite of tools to tackle an often overlooked question: how to deal with sequences from untargeted or unexpected organisms (*i.e. *contaminants but also organelles, commensal or parasitic organisms) in genomic data?*

Availability: *The tools are written in Python3 and R under the GPLv3 Licence and can be found at* https://github.com/itsmeludo/Phyloligo/.

Keywords multi-species, contaminant filtering, untargeted sequencing, oligonucleotide signature, metagenomics

1 Introduction

We present PhylOligo, a new package including tools to explore, identify and extract organism-specific sequences in a genome assembly using the analysis of their DNA compositional characteristics: the oligonucleotide signature.

Compared to existing software, PhylOligo provides several features to explore assemblies including: (i) a customisable oligonucleotide pattern (ii) an interactive cladogram-based visualisation of the contig signature similarity and cumulative size to explore the signature clusters and profile putative additional materials (iii) a species-specific sequence filtering based on a supervised learning of candidate profiles and a double threshold scan.

Our strategy includes 3 main steps:
a) **Assembly exploration** using an **interactive tree visualisation** based on oligonucleotide profiles computed from all genomic contigs. PhylOligo allows for a visual exploration of the compositional similarity distribution and structure of the contigs in an assembly based on either continuous (kmers) or spaced-pattern oligonucleotide frequencies. The oligonucleotide profile of each contig is computed and a pairwise distance matrix based is produced (Figure 1A) to generate an interactive Neighbour-Joining tree. Branch width is drawn proportional to the cumulated length of the contigs in a clade, allowing the user to track where the main part of the assembly clusters (assumed to correspond to the targeted organisms) and what significant clades branch out as hint for separate organisms (see Figure 1B). Thanks to the Ape package, sequences from a clade are interactively selected on the tree and exported to learn a prototype of their oligonucleotide profile.

b) Oligonucleotide **profile prototype learning** based on contig subsets selected by the user at nodes of the tree. ContaLocate then allows the learning of oligonucleotide profiles of the main and presumed additional organisms identified and sampled by the user at the previous step.

c) **Assembly partitioning** to locate organism-specific regions and classify contigs or segments according to the learned prototypes. The assembly is then scanned with sliding windows to locate organism-specific regions using oligonucleotide divergences computed against the targeted and the additional profiles. The distribution of the divergence against both is used to establish two thresholds best separating the different modes in the density

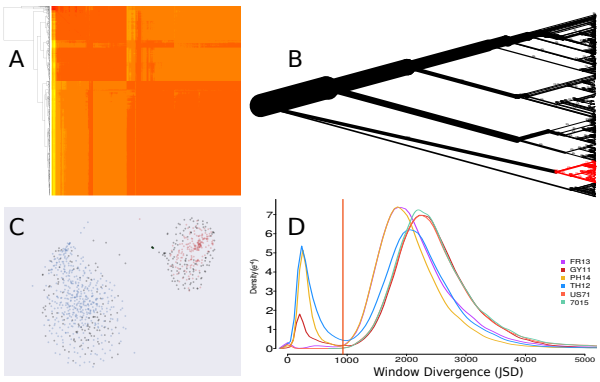


Fig. 1. Visualisation and interactive exploration of assemblies. **A**: Pairwise compositional divergence of contigs produced by PhylOligo. Contigs are reordered by hierarchical clustering. **B**: Contig tree produced by PhylOligo on the tardigrade genome. The clade in red is the current selection pointed by the user. **C**: Contigs clustered by HDBSCAN on oligonucleotide frequencies, Data from *Magnaporthe oryzae*. Red and blue are predicted clusters, grey are unclassified. The hyperspace is reduced to 2 dimensions with t-SNE. **D**: Determination of the untargeted threshold in ContaLocate based on the distribution of distances between the untargeted clade and the scanning windows over the whole assembly.

functions (See Figure 1D). Genomic regions with a divergence simultaneously crossing respective thresholds to the targeted and to the additional profiles are labelled as part of the additional organism and exported as a GFF file.

2 Results

Our strategy present several advantages. 1) Unlike sequence homology based methods, PhylOligo allows the identification of putative uncharacterised and distantly related sequences in assemblies. 2) The double threshold species-specific filtration prevents the removal of HGTs and the subsequent fragmentation of the assembly. 3) Chimeric contigs are detected and split. 4) Compared to filtering unassembled reads, learning the compositional profile on contigs allows for a refined profile, thus leading to a finer filtering.

PhylOligo has been successfully applied to identify untargeted bacterial organisms in four fungi genomic datasets. The software also identified additional organisms in the scaffolds of the tardigrade assembly.

Acknowledgements

We thank the INRA bioinformatics platforms MIGALE and GenoToul for resources and Thomas Schiex and Natalie Villa-Vialaneix for their input.

Dynamix: Dynamic visualization by automatic selection of informative tracks from hundreds of genomic data sets

Matthias MONFORT[1], Eileen E.M. FURLONG[1] and Charles GIRARDOT[1]

[1] European Molecular Biology Laboratory, Genome Biology Unit, Meyerhofstr. 1, D-69117, Heidelberg, Germany

Corresponding Author: matthias.monfort@embl.de

Visualization of genomic data is fundamental for gaining insights into genome function. Yet, co-visualization of a large number of data sets remains a challenge in all popular genome browsers and the development of new visualization methods is needed to improve the usability and user experience of genome browsers.

The Furlong Laboratory presents Dynamix [1], a genome browser plugin for JBrowse [2] that enables the parallel inspection of hundreds of genomic data sets. Dynamix takes advantage of a priori knowledge to automatically display data tracks with signal within a genomic region of interest. As the user navigates through the genome, Dynamix automatically updates data tracks and limits all manual operations otherwise needed to adjust the data visible on screen. Dynamix also introduces a new carousel view that optimizes screen utilization by enabling users to independently scroll through groups of tracks.

Dynamix can be experimented navigating the 4CBrowser (http://furlonglab.embl.de/4CBrowser), a companion website released with Ghavi-Helm et al. 2014 [3], and the DynamixDemo server that demonstrates how Dynamix can be used for rich displays (http://furlonglab.embl.de/DynamixDemo). Visualization of genomics data sets remains a challenging area of research, we hope that Dynamix will encourage the development of innovative visualization methods.

Dynamix is hosted at http://furlonglab.embl.de/Dynamix under the MIT licence and has been published in *Bioinformatics* [1].

References

[1] Matthias Monfort, Eileen E. M. Furlong, Charles Girardot; Dynamix: dynamic visualization by automatic selection of informative tracks from hundreds of genomic datasets. *Bioinformatics*, btx141, 2017

[2] Oscar Westesson, Mitchell Skinner, Ian Holmes, Visualizing next-generation sequencing data with JBrowse. *Brief Bioinform*, 14 (2): 172-177, 2013

[3] Yad Ghavi-Helm, Felix A. Klein, Tibor Pakozdi, Lucia Ciglar, Daan Noordermeer, Wolfgang Huber, Eileen E. M. Furlong, Enhancer loops appear stable during development and are associated with paused polymerase. *Nature*, 512(7512):96-100, 2014

GeneSpy, a simple tool to explore genomic context

Pierre Simon GARCIA[1,2], Frédéric JAUFFRIT[1,3], Christophe GRANGEASSE[2] and Céline BROCHIER-ARMANET[1]

[1] LBBE Laboratoire de Biométrie et Biologie Evolutive, UCB Lyon 1 - Bât. Grégor Mendel 43 bd du 11 novembre 1918, 69622 VILLEURBANNE cedex, France

[2] MMSB Molecular Microbiology and Structural Biochemistry, Institut de Biologie et de Chimie des Protéines 7 passage du Vercors, 69367 LYON7 cedex, France

[3] R&D microbiology, Clinical Unit, bioMerieux, 376 Chemin de l'Orme, 69280 Marcy L'Etoile, France

Corresponding Author: pierre.garcia@etu.univ-lyon1.fr

Genomic context exploration is an important approach to provide informations in phylogenetic and genomic analysis : it brings elements to highlight evolutionary events such as duplications or horizontal transfers, functional links between genes and to assist in identification of orthologs. There are currently some available tools to visualize the genomic context of genes, such as MgcV [1], but they have some drawbacks. Indeed, most of these tools are based on non-local databases (generally made up of a set of complete genomes), do not allow dynamic navigation along genomes, provide non-publishable figures with a restricted selection of export formats (mainly pdf) and are generally slow to generate figures. In this context, we developed an application able to quickly generate visualizations of genomic context of genes, GeneSpy. This tool allows users to explore the organization of genomes and compare them through a user-friendly interface.

GeneSpy is written in python 2.7 and depends on Matplotlib and Tkinter libraries. GeneSpy uses gff files to build its local database. A download utility can be used to retrieve and prepare gff files for a set of assemblies from genbank or refseq. The list of target genes can be provided by the user via a tabular text file containing assembly identifiers of genomes and accession identifiers of proteins. The tool also supports BLAST reports from NCBI. Alternatively, users can search genes using keywords (fragment of name of strain and accession number, locus tag, name or annotation).

The context visualization in itself is heavily customizable. Many options are available such as window length, arrow size, type of coloration or verbosity of strains description. By default, the colors used for genes are procedurally generated using protein names and annotations for color consistency between contexts. Manual editing of color attribution is possible through a dedicated menu. The option to color only the genes of interest is also available.

The user can access informations relative to any displayed gene such as accession number, locus tag and predicted function. Moreover, it is possible to update the list of target genes by selecting any gene shown in the context figure. Thus, the user can simply navigate along the genome.

Figures can be exported in many formats such as png, jpg, tif, svg, pdf and multiple pdf. An iTOL-specific export format is also available. It consists of a text file that can be imported in iTOL [2]. This format can be used to annotate a tree with relevant genomic context information, which can be very useful to study evolutionary dynamic of genomic regions. It is worth noting that database can be generated independently of GeneSpy, meaning that a pre-existing local databases can be reused with GeneSpy.

Thus, we provide a tool able to generate publishable figures of genomic context that is simple to use, fast, flexible, and adapted to any database. GeneSpy is available at https://lbbe.univ-lyon1.fr/GeneSpy/ and distributed under CeCILL licence.

This work was supported by grant from ARC1 Santé Rhône-Alpes Auvergne.

References

[1] Overmars L, Kerkhoven R, Siezen RJ, Francke C. MGcV: the microbial genomic context viewer for comparative genome analysis. BMC Genomics; 14:209. 2013.

[2] Letunic I, Bork P. Interactive tree of life (iTOL) v3: an online tool for the display and annotation of phylogenetic and other trees. Nucleic Acids Res.;44(W1):W242-5. 2016.

MCXpress: An R Package for functional interpretation of single cell RNA-Seq data using multivariate analysis.

Akira CORTAL[1], Antonio RAUSELL[2]

1,2Imagine InstituteClinical Bioinformatics Lab, 75015, Paris, France

Corresponding Author: akira.cortal@institutimagine.org
antonio.rausell@institutimagine.org

Abstract

Single-cell RNA-sequencing allows unbiased transcriptome profiling of individual cells, enabling the analysis of genes expression at the cellular level. By uncovering cell heterogeneity in a given cell type, discovery of rare subpopulations of cells which could be responsible of the onset and progression of specific diseases are possible. Here we developed a comprehensible tool for the analysis of RNA-seq data using an approach based on Multiple Correspondence Analysis (MCA). MCA is a dimensionality reduction technique that allows the representation of both individuals (cells) and the variables (genes) within the same Euclidean space, thus allowing the simultaneous identification of subpopulations of cells and their gene signatures. MCA was coupled with gene set enrichment analysis; a powerful analytical method for characterizing differentially expressed genes/pathways within each groups of individuals. This singular combination allows a joint comparison of gene expression and pathway enrichment across all the groups. Combined with a shiny data visualization interface, the MCXpress R package can enhance the interpretation of both single and bulk RNA-sequencing data analysis. The tool development version is available on github at https://github.com/cbl-imagine/MCXpress (MIT License).

References

[1] Shalek, A. K. et al. Single-cell transcriptomics reveals bimodality in expression and splicing in immune cells, Nature 498, 236–240, 2013.
[2] Buettner F et al, Computational analysis of cell-to-cell heterogeneity in single-cell RNA-sequencing data reveals hidden subpopulations of cells, Nature Biotechnology 33(2) 155, 2015.
[3] Rausell A et al, Protein interactions and ligand binding: From protein subfamilies to functional specificity,PNAS 107(5) 1995, 2010.

Deciphering The Functional Effects of Genetic Variation With UniProt Annotations.

Benoit BELY[1], Andrew NIGHTINGALE[1] and Maria MARTIN[1]

[1] Protein Function Development EMBL-EBI, Wellcome Trust Genome Campus, CB10 1SD, Hinxton, UK

Corresponding Author: benoit.bely@ebi.ac.uk

1 Introduction

Variants such as, substitution, frame shift and in frame deletions and insertions, in protein coding genes may cause deleterious effects manifested as debilitating inherited diseases or syndromes. Variants that cause a disease, do so by altering gene transcription or key parts of the proteins resulting in the disruption of a protein's native structure or its ability to execute functions, for example, to catalyze chemical reactions. The UniProt KnowledgeBase (UniProtKB) provides protein function information including annotation of protein altering variants, with any known functional effects and disease associations taken from scientific literature and large-scale public experimental data sets. Providing this rich knowledge at the genetic level is essential for discovering the potential deleteriousness of genetic variants.

2 Human proteome variation

UniProt has mapped protein annotations in the human proteome to the GRCh38 assembly of the human genome. Twenty-seven structural and functional annotations are currently provided including: enzyme active sites, modified residues, protein binding domains, protein variations etc. The complete set of human proteome sequences, including isoforms are also provided, for comparison to predicted transcripts. These mappings and related annotations have been made available as text BED files and BigBed files that are compatible with most genome browsers (http://www.uniprot.org/downloads). These files are also bundled into a public track hub that is available with the Ensembl and UCSC genome browsers through their public track hub registries or can be accessed directly from the genome track hubs registry (https://trackhubregistry.org). UniProt protein annotations can also be integrated directly into large scale genomic data programmatically via the Proteins API (http://www.ebi.ac.uk/proteins/api/doc/) [1].

3 Disease associated variants

To illustrate how the integration of protein function knowledge with high throughput genomic data provides unique opportunities for biomedical research. We examine some specific biological examples in disease related genes and proteins illustrating the utility of the combining protein and genome annotations for the functional interpretation of variants. Thus, showing that the UniProt genomic mappings, can help scientists to rapidly comprehend complex processes in biology.

References

[1] Nightingale A., Antunes R., Alpi E., Bursteinas B., Gonzales L., Liu W., Luo J., Qi G., Turner E. and Martin, M. The Proteins API: accessing key integrated protein and genome information. *NAR Webservices Edition*, 2017.

Conférence invitée

John Huelsenbeck

Department of Integrative Biology, UC Berkeley, USA

Bayesian inference in phylogeny for genome-scale data

Bayesian inference has permeated the field of phylogenetics. A major challenge in the field remains how to extend methods to genome-scale data. The temptation is to take short-cuts by applying fast methods that do not take full advantage of the information contained in the data. I describe several methods that may be applicable to genome-scale data. First, I describe new proposal mechanisms for better inferring large phylogenetic trees. Second, I discuss a class of models that can be used to address questions such as the identification of sites under the influence of natural selection.

Extreme halophilic archaea derive from two distinct methanogen Class II ancestors

Monique Aouad[1], Najwa Taib[1], Anne Oudart[1], Michel Lecocq[1], Manolo Gouy[1] and Céline Brochier-Armanet[1]

[1] Univ Lyon, Université Lyon 1, CNRS, UMR5558, Laboratoire de Biométrie et Biologie Évolutive, 43 bd du 11 novembre 1918, F-69622, Villeurbanne, France

Corresponding Author: monique.aouad@univ-lyon1.fr

Abstract *The phylogenetic analysis of conserved core genes has disentangled most of the ancient relationships in Archaea. However, some groups remain debated, like the DPANN, a recently proposed deep-branching super-phylum gathering various lineages of nanosized archaea with reduced genomes. Among these, the Nanohaloarchaea thrive in high-salt environments and require high-salt concentrations for growth. The discovery of Nanohaloarchaea in 2012 was significant because at the time extreme halophilic archaea were represented by a single lineage, the Halobacteria, a major class belonging to Euryarchaeota. The phylogenetic position of Nanohaloarchaea is highly debated, being alternatively proposed as the sister-lineage of Halobacteria or member of the DPANN super-phylum. Pinpointing the phylogenetic position of extreme halophilic archaea is important to improve our knowledge of the deep evolutionary history of Archaea and decipher the underlying molecular adaptive processes and the evolutionary paths that allowed their emergence. Using comparative genomic approaches, we identified more than 250 protein markers carrying a reliable phylogenetic signal to address this issue. By combining strategies limiting the impact of biases on tree inferences, we showed that Nanohaloarchaea and Halobacteria represent two independent lineages that derived from methanogens Class II. This implies that adaption to very high salinity emerged twice independently in Archaea and that the grouping of Nanohaloarchaea within DPANN lineages is the consequence of a tree reconstruction artifact, which could challenge the existence of this group.*

Keywords Stenosarchaea, compositional bias, long branch attraction, Slow-Fast method, rate signal

Introduction

Recent advances in high-throughput sequencing technologies have revealed many new major uncultured environmental groups, most of them being known only through ribosomal RNA or genomic sequences [1–3]. This is for instance the case of *Nanohaloarchaea*, a group of extreme halophilic nanosized *archaea*, discovered recently in Lake Tyrell, Australia [4]. Some studies suggested that they represent the sister-lineage of *Halobacteria* [4,5], the other lineage of extreme halophilic archaea, while other analyses, based on different sets of markers, different methods and/or different taxonomic samplings, suggested instead that *Nanohaloarchaea* belong to the recently proposed DPANN super-phylum. This deep-branching group encompasses diverse fast evolving, possibly nanosized archaea (e.g. *Diapherotrites*, *Parvarchaeota*, *Micrarchaeota*, *Aenigmarchaeota*, *Nanoarchaeota*, *Woesearchaeota*, *Pacearchaeota*) including *Nanohaloarchaea* [2,3,6]. Regarding *Halobacteria*, phylogenetic analyses of the RNA component of the small subunit of the ribosome and large supermatrices of conserved core genes have revealed a close relationship between *Halobacteria* and methanogens Class II, a group encompassing *Methanomicrobiales*, *Methanosarcinales* and *Methanocellales*, and indicated that *Halobacteria* could derive from a methanogenic ancestor [7]. However, the identity of the closest relative of *Halobacteria* remains debated. In fact, recently published phylogenies supported *Halobacteria* as the sister-lineage of all methanogens Class II [8–11], of *Methanomicrobiales* [12–14], or of *Methanocellales* [5,15].

Elucidating the precise position of *Halobacteria* and *Nanohaloarchaea* is particularly challenging because their proteomes harbor atypical amino acid compositions as a consequence of their extremophilic lifestyle. This can generate a compositional signal that may conflict with and dominate over the phylogenetic signal

[16], and lead to artifactual tree reconstructions where distant sequences with similar compositions are clustered together [17,18]. Another source of bias could be linked to the fast evolutionary rate of nanohaloarchaeal and halobacterial proteomes highlighted by their very long branches in phylogenetic trees compared to other archaeal lineages [4,5]. The phylogenetic position of fast-evolving species and long branches is particularly difficult to determine because differences in evolutionary rates among lineages can generate a rate signal that may conflict the phylogenetic signal [16] and cause tree reconstruction artifacts such as the long branch attraction (LBA) [19]. This well-known tree reconstruction artifact tends to group fast-evolving sequences/long branches and slow-evolving sequences/short branches in different parts of the trees when the rate signal dominates over the phylogenetic signal [16]. Accordingly, we may wonder to what extent the conflicting positions observed for *Nanohaloarchaea* and *Halobacteria* are the consequence of tree reconstruction artifacts and if it is possible to overcome them.

To address this issue we performed an in-depth phylogenomic analysis designed to limit the impact of the non-phylogenetic signal on phylogenetic inferences. We showed that *Nanohaloarchaea* and *Halobacteria* branch robustly with *Methanocellales* and *Methanomicrobiales*, respectively, meaning that they derive from two distinct methanogen Class II ancestors. This implies also that adaptation to very high salinity occurred at least twice in archaea, and that the phenotypical similarities of *Nanohaloarchaea* and *Halobacteria* result from convergent evolutionary processes, possibly accompanied by horizontal gene transfers. Finally, our results indicate also that the grouping of *Nanohaloarchaea* with other DPANN lineages is the consequence of a tree reconstruction artifact, challenging the existence of this candidate super-phylum.

Results

The comparison of 155 proteomes from ANME-1, methanogens Class II, *Halobacteria* and their close relatives: *Archaeoglobales* and *Diaforarchaea*, and from *Nanohaloarchaea* led to the delineation of 108,007 families, among which 258 presented a broad taxonomic distribution and no or very few evidences of horizontal gene transfers (HGT) among these lineages and/or gene duplications. *Diaforarchaea* represent the first diverging lineage, according to previous studies.

To test the impact of missing data on phylogenetic inference, different versions of these supermatrices were built by gathering protein families present in more than 95% or 70% of the studied proteomes. The 68 protein markers present in the three *Nanohaloarchaea* and in more than 70% of the 155 studied proteomes (including *Halobacteria*) were combined to build the FNANOHALO$_{70}$ supermatrix. The corresponding ML tree was overall well resolved (Fig 1). In particular, the sistership between *Methanocellales* and *Methanosarcinales* was recovered (BV = 100%), indicating that adding *Nanohaloarchaea* and *Halobacteria* did not cause major tree reconstruction artifacts. Regarding extreme halophiles, *Nanohaloarchaea* and *Halobacteria* did not group together. In fact, *Halobacteria* clustered with *Methanomicrobiales* (BV = 99%, Fig 1), in agreement with recent studies [12–14], whereas *Nanohaloarchaea* branched on the stem of *Diaforarchaea* (BV = 99%, Fig 1). This indicated that *Nanohaloarchaea* are not related to any methanogens Class II lineage, ANME-I or *Archaeoglobales*, and is compatible with a deep-branching position of *Nanohaloarchaea* within *Archaea*, as postulated by the DPANN hypothesis [2,3,6].

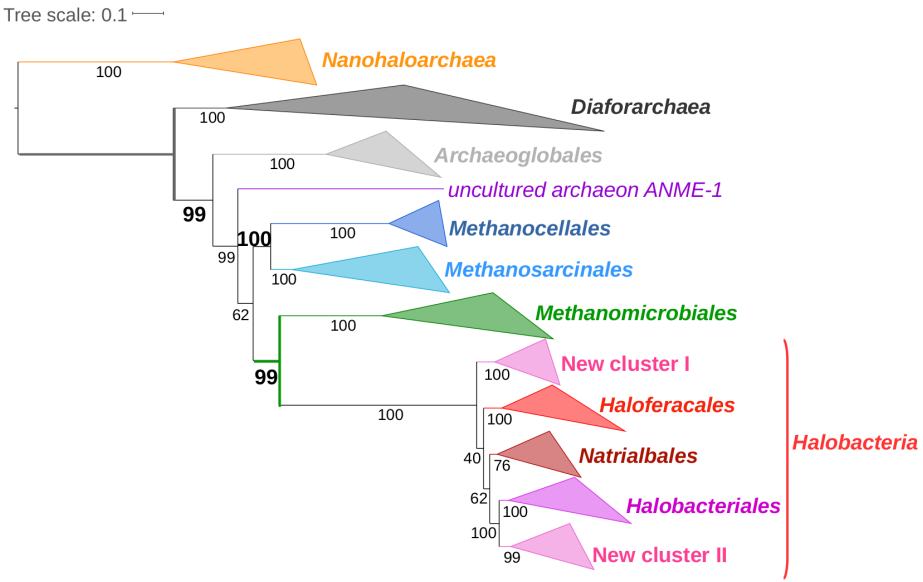


Fig 1. Maximum Likelihood phylogeny inferred with the FNANOHALO$_{70}$ supermatrix (18,309 positions, 155 sequences). The tree was inferred with PHYML 3.1 using the LG+I+G4+F model as proposed by IQ-TREE (BIC criteria). The scale bar corresponds to the average number of substitutions per site. Numbers at nodes correspond to bootstrap supports (100 replicates of the original dataset).

Sequence compositional heterogeneity is an important source of bias in molecular phylogeny [16,20]. To test the impact of compositional biases on the inferred phylogenies, we removed from the supermatrices the positions most strongly responsible for the amino acid composition heterogeneity among sequences. The removal of positions with the highest compositional bias did not impact the phylogenetic position of extreme halophilic archaea. The clustering of *Halobacteria* with *Methanomicrobiales* and the branching of *Nanohaloarchaea* on the stem of *Diaforarchaea* were recovered again with high supports.

The evolutionary rate signal is one of the major causes of tree reconstruction artifacts such as the LBA [17,19]. This artifact is caused by multiple substitutions occurring at the same sites, a process which erases progressively the most ancient phylogenetic signal and results in the grouping of sequences according to their evolutionary rates (i.e. rate signal) in different parts of the inferred trees. To overcome this issue, dedicated methods have been developed.

Among them, the recoding of amino acids allows to hide substitutions among similar amino acids. To test the impact of the rate signal, we applied two different recoding schemes (dayhoff4 and dayhoff6) to all supermatrices in a Bayesian framework. The inferred BI trees confirmed the sister-ship between *Halobacteria* and *Methanomicrobiales*. Yet, surprisingly, *Nanohaloarchaea* branched with *Methanocellales* in seven out of the eight recoded supermatrices (Fig 2), suggesting that the branching of *Nanohaloarchaea* on the stem of *Diaforarchaea* could result from a tree reconstruction artefact due to the rate signal.

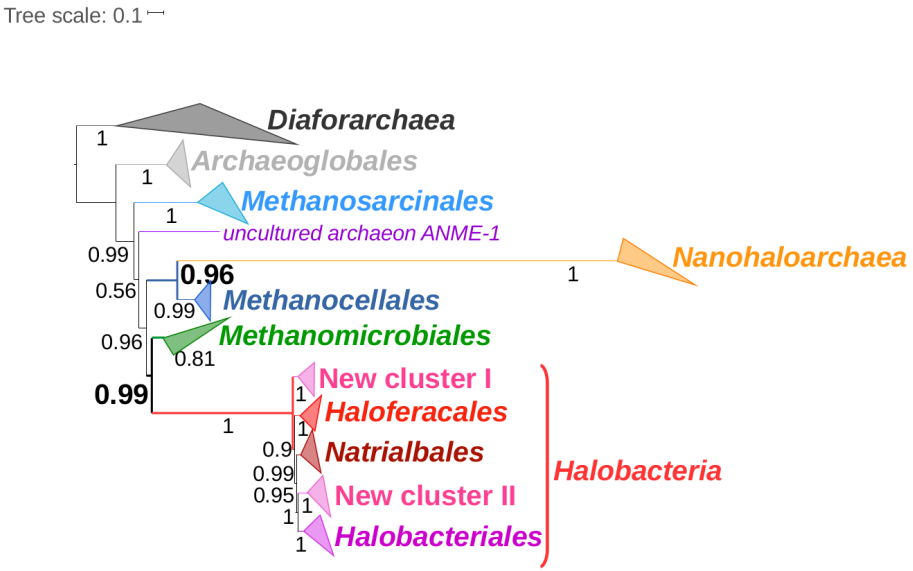


Fig 2. Bayesian phylogeny inferred with the FNANOHALO$_{70}$ supermatrix recoded according to the Dayhoff4 scheme (18,309 positions, 155 sequences). The tree was inferred with Phylobayes using the CAT + GTR + G4 evolutionary model. The scale bar corresponds to the average number of substitutions per site. Numbers at nodes correspond to posterior probabilities.

To test this hypothesis, we used another approach to limit tree reconstruction artifacts resulting from the rate signal. This approach, called the Slow-Fast method (S-F), consists in the progressive removal of the fastest evolving sites from multiple alignments [21]. The S-F method was shown to be very efficient to reduce tree reconstruction artifacts because the fastest evolving sites are the most susceptible to be impacted by multiple substitutions [17,22]. This approach allows monitoring the support associated to a given branch of a tree throughout the removal process and thus to determine if the corresponding relationship reflects the phylogenetic or the rate signal contained in sequences [17]. To avoid biases in the estimation of evolutionary rates at each position due to missing data and/or unbalanced taxonomic sampling among lineages, the S-F method was applied to the supermatrices built with markers present in at least 95% of the studied taxa and by keeping only three to seven representative sequences for each archaeal lineage.

Removal of the fastest-evolving sites did not impact the phylogenetic position of *Halobacteria*. In fact, the grouping of *Halobacteria* with *Methanomicrobiales* was strongly supported in ML and BI trees inferred with the S-F supermatrices. This suggested that this relationship was not the consequence of the rate signal. In sharp contrast, the S-F method showed that a robust grouping of *Nanohaloarchaea* with *Methanocellales* was observed in ML and BI trees when the fastest-evolving sites were removed (Fig 3).

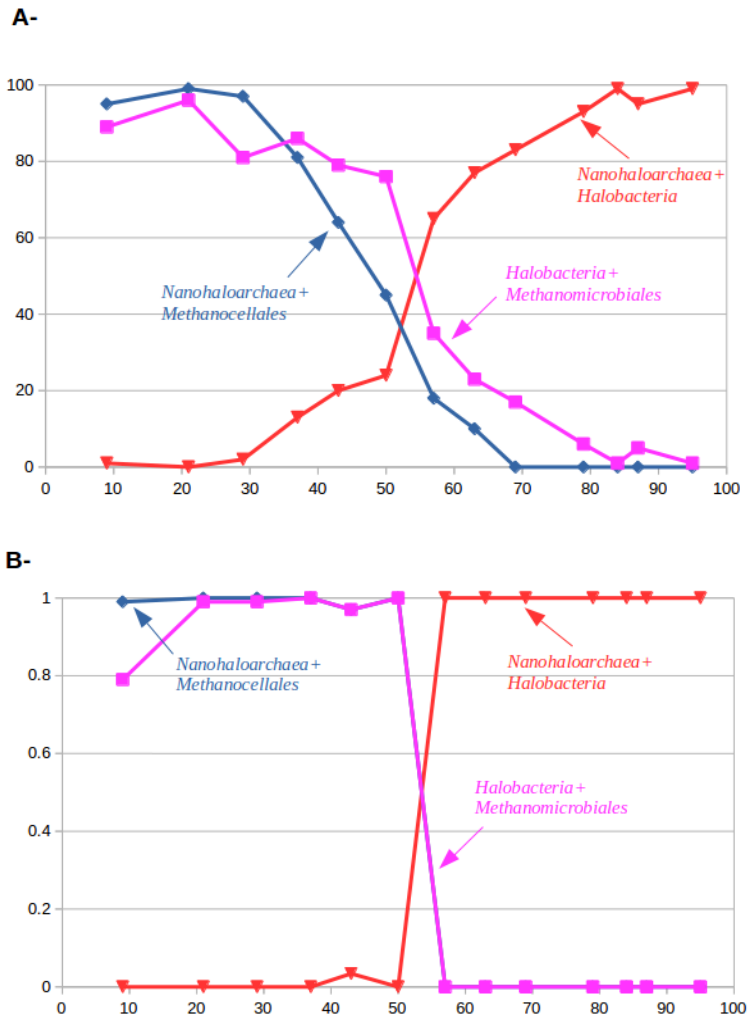


Fig 3. Effect of removal of fast-evolving positions on the phylogenetic position of *Halobacteria* and *Nanohaloarchaea* in the FNANOHALO$_{95}$ S-F matrices. The x axes correspond to the fraction of sites (in %) kept in the supermatrices. The removal of fastest-evolving sites proceeds from right to left. The y axes indicate the support for the grouping of *Halobacteria* with *Methanomicrobiales* (in pink), the grouping of *Nanohaloarchaea* with *Halobacteria* (in red), and the grouping of *Nanohaloarchaea* with *Methanocellales* (in blue) in ML (A) and BI trees (B) resulting from the S-F analysis. Notice that the *Halobacteria* + *Nanohaloarchaea* grouping supported by FNANOHALO$_{95}$ disappears when fast-evolving site are removed.

Discussion

By focusing our analysis on the part of the euryarchaeal tree that contains *Halobacteria* and methanogens Class II, we were able to assemble larger datasets of conserved markers and to use more intense taxonomic samplings compared to previous studies. More precisely, we identified 258 conserved protein families widely distributed in the 155 taxa, among which 68 were present in the three representatives of *Nanohaloarchaea* available at the time. This led us to the construction of large supermatrices containing thousands of amino acid positions. By using various methods allowing to decouple the different types of signal contained in protein sequences, we showed that the compositional signal did not significantly impact the phylogenetic positions of *Nanohaloarchaea* and *Halobacteria*, while the rate signal had a major impact on the phylogenetic position of *Nanohaloarchaea*. In fact, two independent methods allowing to reduce the impact of multiple substitutions on phylogenetic inferences, the removal of the fastest evolving sites and the recoding of amino acids, provided consistent results supporting the grouping of *Halobacteria* with *Methanomicrobiales* and of *Nanohaloarchaea* with *Methanocellales.* The robust and recurrent grouping of *Halobacteria* and methanogens Class II in many studies, suggested that they could represent a new super-class, that we propose to call *Stenosarchaea* (from the Greek stenós, meaning close/joint). Regarding *Nanohaloarchaea*, our analyses strongly suggested that they are part of the *Stenosarchaea*, and more precisely, that they represent the sister group to *Methanocellales*.

The grouping of *Nanohaloarchaea* with *Methanocellales* and that of *Halobacteria* with *Methanomicrobiales* within *Stenosarchaea* has major implications and opens new perspectives. First, it implies that both lineages derive from two distinct but related methanogen ancestors, which is in accordance with the fact that a few archaeal lineages, all belonging to *Euryarchaeota*, can survive at high salt concentrations [23]. These halophilic or salt-tolerant archaea are anaerobic methanogens living in hypersaline sediments. Most are methylotroph and belong to methanogens Class II and more precisely to *Methanosarcinaceae* (order *Methanosarcinales*) or *Methanocalculaceae* (order *Methanomicrobiales*) [23]. It also implies that adaptation to extreme high salt concentrations occurred at least twice independently during the evolution of *Archaea*. Thus, the phenotypic properties shared by *Nanohaloarchaea* and *Halobacteria* should be interpreted as the consequence of a convergent evolution that could have been facilitated by HGT. In that context, it would be interesting to reevaluate the evolutionary history of these lineages, and the role played by HGT in the emergence of *Halobacteria* and *Nanohaloarchaea* from methanogenic ancestors.

Finally, the robust grouping of *Nanohaloarchaea* with *Methanocellales* rules out alternative hypotheses for the branching of *Nanohaloarchaea*, and in particular a branching outside of *Stenosarchaea*, as expected if *Nanohaloarchaea* were part of the candidate DPANN superphylum. This challenges the existence of this group as it is currently described and questions to what extent similar artifacts could also impact the position of the other lineages composing this super-phylum, which are all fast evolving. Testing this hypothesis would require accurate and separate analyses, each focused on one lineage.

References

1. Schleper C, Jurgens G, Jonuscheit M. Genomic studies of uncultivated archaea. Nat. Rev. Microbiol. [Internet]. 2005 [cited 2016 Jun 20];3:479–88. Available from: http://www.ncbi.nlm.nih.gov/pubmed/15931166

2. Rinke C, Schwientek P, Sczyrba A, Ivanova NN, Anderson IJ, Cheng J-F, et al. Insights into the phylogeny and coding potential of microbial dark matter. Nature [Internet]. 2013 [cited 2016 Jun 20];499:431–7. Available from: http://www.ncbi.nlm.nih.gov/pubmed/23851394

3. Castelle CJ, Wrighton KC, Thomas BC, Hug LA, Brown CT, Wilkins MJ, et al. Genomic expansion of domain archaea highlights roles for organisms from new phyla in anaerobic carbon cycling. Curr. Biol. [Internet]. 2015 [cited 2016 Jun 20];25:690–701. Available from: http://www.ncbi.nlm.nih.gov/pubmed/25702576

4. Narasingarao P, Podell S, Ugalde JA, Brochier-Armanet C, Emerson JB, Brocks JJ, et al. De novo metagenomic assembly reveals abundant novel major lineage of Archaea in hypersaline microbial communities. ISME J. [Internet]. 2012 [cited 2016 Jun 20];6:81–93. Available from: http://www.ncbi.nlm.nih.gov/pubmed/21716304

5. Petitjean C, Deschamps P, López-Garciá P, Moreira D. Rooting the domain archaea by phylogenomic analysis supports the foundation of the new kingdom Proteoarchaeota. Genome Biol. Evol. 2014;7:191–204.

6. Williams TA, Embley TM. Archaeal "dark matter" and the origin of eukaryotes. Genome Biol. Evol. [Internet]. 2014 [cited 2016 Jun 20];6:474–81. Available from: http://www.ncbi.nlm.nih.gov/pubmed/24532674

7. Forterre P, Brochier C, Philippe H. Evolution of the Archaea. Theor. Popul. Biol. [Internet]. 2002 [cited 2016 Jun 20];61:409–22. Available from: http://www.ncbi.nlm.nih.gov/pubmed/12167361

8. Gao B, Gupta RS, Woese C, Kandler O, Wheelis M, Pace N, et al. Phylogenomic analysis of proteins that are distinctive of Archaea and its main subgroups and the origin of methanogenesis. BMC Genomics [Internet]. BioMed Central; 2007 [cited 2016 Jun 20];8:86. Available from: http://bmcgenomics.biomedcentral.com/articles/10.1186/1471-2164-8-86

9. Yutin N, Puigbò P, Koonin E V., Wolf YI, Ramakrishnan V, Ban N, et al. Phylogenomics of Prokaryotic Ribosomal Proteins. Lespinet O, editor. PLoS One [Internet]. Public Library of Science; 2012 [cited 2016 Jun 20];7:e36972. Available from: http://dx.plos.org/10.1371/journal.pone.0036972

10. Wolf YI, Makarova KS, Yutin N, Koonin E V. Updated clusters of orthologous genes for Archaea: a complex ancestor of the Archaea and the byways of horizontal gene transfer. Biol. Direct [Internet]. 2012 [cited 2016 Jun 20];7:46. Available from: http://www.ncbi.nlm.nih.gov/pubmed/23241446

11. Nelson-Sathi S, Dagan T, Landan G, Janssen A, Steel M, McInerney JO, et al. Acquisition of 1,000 eubacterial genes physiologically transformed a methanogen at the origin of Haloarchaea. Proc. Natl. Acad. Sci. U. S. A. [Internet]. 2012;109:20537–42. Available from: http://www.pnas.org/content/109/50/20537.abstract

12. Brochier-Armanet C, Forterre P, Gribaldo S. Phylogeny and evolution of the Archaea: One hundred genomes later. Curr. Opin. Microbiol. 2011. p. 274–81.

13. Raymann K, Brochier-Armanet C, Gribaldo S. The two-domain tree of life is linked to a new root for the Archaea. Proc. Natl. Acad. Sci. U. S. A. [Internet]. 2015 [cited 2016 Jun 20];112:6670–5. Available from: http://www.ncbi.nlm.nih.gov/pubmed/25964353

14. Petitjean C, Deschamps P, López-García P, Moreira D, Brochier-Armanet C. Extending the conserved phylogenetic core of archaea disentangles the evolution of the third domain of life. Mol. Biol. Evol. 2015;32:1242–54.

15. Becker EA, Seitzer PM, Tritt A, Larsen D, Krusor M, Yao AI, et al. Phylogenetically Driven Sequencing of Extremely Halophilic Archaea Reveals Strategies for Static and Dynamic Osmo-response. Whitaker RJ, editor. PLoS Genet. [Internet]. Public Library of Science; 2014 [cited 2016 Jun 20];10:e1004784. Available from: http://dx.plos.org/10.1371/journal.pgen.1004784

16. Jeffroy O, Brinkmann H, Delsuc F, Philippe H. Phylogenomics: the beginning of incongruence? Trends Genet. 2006;22:225–31.

17. Delsuc F, Brinkmann H, Philippe H. Phylogenomics and the reconstruction of the tree of life. Nat. Rev. Genet. [Internet]. Nature Publishing Group; 2005 [cited 2016 Jun 20];6:361–75. Available from: http://www.nature.com/doifinder/10.1038/nrg1603

18. Woese CR, Achenbach L, Rouviere P, Mandelco L. Archaeal phylogeny: reexamination of the phylogenetic position of Archaeoglobus fulgidus in light of certain composition-induced artifacts. Syst. Appl. Microbiol. [Internet]. 1991 [cited 2016 Jun 20];14:364–71. Available from: http://www.ncbi.nlm.nih.gov/pubmed/11540072

19. Felsenstein J. Cases in which Parsimony or Compatibility Methods Will be Positively Misleading. Syst. Zool. [Internet]. 1978 [cited 2016 Jun 20];27:401. Available from: http://www.jstor.org/stable/2412923?origin=crossref

20. Gribaldo S, Philippe H. Ancient Phylogenetic Relationships. Theor. Popul. Biol. [Internet]. 2002;61:391–408. Available from: http://www.scopus.com/inward/record.url?eid=2-s2.0-0036593395&partnerID=40&md5=c809464d9ca50344aef3f1c0ecd9c871

21. Brinkmann H, Philippe H. Archaea sister group of Bacteria? Indications from tree reconstruction artifacts in ancient phylogenies. Mol. Biol. Evol. [Internet]. 1999 [cited 2016 Jun 20];16:817–25. Available from: http://www.ncbi.nlm.nih.gov/pubmed/10368959

22. Philippe H. Opinion: long branch attraction and protist phylogeny. Protist [Internet]. 2000 [cited 2016 Jun 20];151:307–16. Available from: http://www.ncbi.nlm.nih.gov/pubmed/11212891

23. Oren A. Taxonomy of halophilic Archaea: current status and future challenges. Extremophiles [Internet]. 2014 [cited 2016 Jun 20];18:825–34. Available from: http://www.ncbi.nlm.nih.gov/pubmed/25102811

Origin and evolution of multiple haem copper oxidases in *Archaea*

Anne Oudart[1], Simonetta Gribaldo[2] and Céline Brochier-Armanet[1]

[1]Univ Lyon, Université Lyon 1, CNRS, UMR5558, Laboratoire de Biométrie et Biologie Évolutive, 43 bd du 11 novembre 1918, F-69622, Villeurbanne, France.

[2]Unité de Biologie Moléculaire du Gene chez les Extremophiles, Département de Microbiologie, Institut Pasteur, 75015 Paris, France

Corresponding Author: celine.brochier-armanet@univ-lyon1.fr

Abstract

Understanding the origin and evolution of dioxygen reductases, the terminal electron acceptors of aerobic respiratory chains, can provide precious clues for the emergence of this energy process that is still debated. The haem-copper oxidases superfamily (HCO), belonging to the dioxygen reductases, contains three families of dioxygen reductases named A, B, and C according to a classification based on sequence similarity and phylogenetic analysis of the homologous catalytic subunits [1]. A phylogenomic study showed that these enzymes have very different evolutionary histories [2] with an ancient dioxygen reductase (A-HCO) present prior to the divergence of major present-day bacterial and archaeal phyla, thus before the emergence of oxygenic photosynthesis. However, this result is in contradiction with those proposed by Ducluzeau et al. 2014 [3] about the structure of the A-HCO suggesting that this dioxygen reductase would be the most recent. So the question about the origin of the haem-copper oxidases is still unresolved, and a new analysis is required. The haem-copper oxidases superfamily of the ancestor of Archaea probably used dioxygen [4]. Nevertheless, the previous conclusions are old and the available data was very limited (73 archeal complete genomes in 2009). Today, with more available data in the public database (252 archeal complete genomes) it is interesting to reassess the questions about the origin and evolution of the haem-copper superfamily for Archaea. We will present our results about subunits of the haem-copper superfamily in Archaea.

Keywords Aerobic respiration, LUCA, Phylogeny, Prokaryotes, Early earth

Introduction

Dioxygen reductases (O_2-Red) represent the terminal electron-transfer enzymes of aerobic respiratory chains in the three Domains of Life (*Archaea*, *Bacteria* and *Eucarya*). These membrane-bound enzymes catalyze the reduction of dioxygen (O_2) to water by using electrons provided by either a quinol derivate or a cytochrome *c*. Because O_2-Red are key enzymes of aerobic respiration, understanding their origin and evolution could provide precious clues for the emergence of this key metabolism, the rise of O_2 on early Earth and its impact on the evolution of ancient microbial communities. The O_2-Red encompasses HCO divided into three subfamilies named A, B, and C according to a recently proposed classification based on sequence similarity and phylogenetic analyses of their catalytic subunits [1,5]. The O_2-Red include also the Nitric Oxide Reductases (NOR) which reduce nitric oxide (NO) to nitrous oxide (N2O) [2].

The catalytic subunit of A-HCO is found in many bacterial lineages [2]. In archaea, it was detected in *Crenarchaeota* and Thaumarchaeota (two phyla belonging to the TACK super-phylum), and in *Diaforarchaea* and *Halobacteria*, two euryarchaeotal orders [2]. The catalytic subunit of B-HCO was reported in a few bacterial lineages, in *Crenarchaeota* and in *Halobacteria [2]*. The catalytic subunit of C-HCO is mainly present in *Proteobacteria* [2]. Finaly NOR catalytic subunit was mainly found in *Proteobacteria*, in some other *Bacteria*, and in a few *Archaea* [2].

The evolutionary history of HCO O_2-Red is highly debated, and three different scenarii have been proposed. First, it was proposed that the four HCO O_2-Red were present in the Last Universal Common Ances-

tor (LUCA) [6]. More recently, it was proposed that only A-HCO was present in the ancestor of Bacteria, and possibly in LUCA, while the three other HCO O_2-Red subfamilies emerged later, in *Archaea* (B-HCO) and in *Proteobacteria* (NOR and C-HCO) [2,4]. According to these two scenarii, the LUCA and/or the ancestor of Bacteria could have been able to reduce O_2. Because these two organisms are ancestors of *Cyanobacteria*, the scenarii implied that some sources of O_2 should have been present before the emergence of oxigenic photosynthesis. In contrast, the third scenario proposed the NOR to be present in the LUCA, while HCO O_2-Red emerged during the diversification of *Bacteria* after the emergence of *Cyanobacteria* and oxigenic photosynthesis, and spread through HGT in archaeal and bacterial lineages [1,3,7]. According to this scenario, HCO O_2-Red acquired their capacity to reduce O_2 independently from distinct NOR ancestors after the emergence of *Cyanobacteria* and of oxygenic photosynthesis.

At the time, the data available were mainly limited to *Bacteria* because the number of genomes (complete or not) of *Archaea* was very restricted compared to *Bacteria*. This lack of data in *Archaea* limited considerably our capacity to understand the ancient evolutionary history of HCO O_2-Red thus of one of the most important metabolism on earth. Fortunately, the situation is changing rapidly, and now hundred genomes of *Archaea* are available in public databases, thanks to ambitious projects aiming at sequencing the cultured and the uncultured "dark matter" of the microbial diversity [8,9]. This rainfall of genomic data offers the possibility to investigate the evolutionary history of HCO O_2-Red in *Archaea*.

In *Archaea*, HCO O_2-Red have been extensively studied in *Sulfolobales*. In these *Crenarchaeota*, A-HCO and B-HCO, the catalytic subunits of O_2-Red, are associated with several proteins to form various different protein complexes (Figure 1).

	A-HCO	B-HCO			
Subunits	**Sox**	**Sox**	**Dox**	**Fox**	**Function**
Subunit I	SoxM	SoxB	DoxB	FoxA/FoxA'	Reduce dioxygen
Subunit II	SoxH	SoxA	DoxC	FoxB	Transfer electron to Subunit I
Subunit III	Can be fused with SoxM				Assist the final step of the folding of the subunit I
Cytochrome b	SoxG	SoxC			Transfer electron to Rieske protein
Cytochrome b558/566				FoxCD	Transfer electron to multi blue copper oxidase
Rieske protein	SoxF	SoxL			Transfer electron to Sulfocyanin
Sulfocyanin	SoxE				Transfer electron to Subunit II
Protein of unknown function	SoxI				Unknown
Protein of unknown function		SoxD			Unknown
Protein of unknown function		SoxD'			Unknown
Protein of unknown function			DoxE		Unknown

Fig 1. Haem copper oxidases complexes in *Sulfolobales*.

The four complexes described in *Archaea* carried homologous subunits, suggesting that these respiratory protein complexes may derive from an ancestral protein complex.

Material and methods

Dataset assembly

The 2,610 bacterial complete proteomes and 391 archaeal complete and incomplete proteomes available at the NCBI (http://www.ncbi.nih.gov) and the JGI (http://jgi.doe.gov/) were retrieved and gathered in a local database. These corresponded to 1279 bacterial and 269 archaeal species.

Identification of the homologues of HCO subunits was performed by combining iterative sequence similarity searches and HMM profiles surveys using the BLASTP v2.2.26 [10] and the HMMER v3.1b1 programs [11] respectively.

Classification of HCO catalytic subunit

HCO catalytic subunits were classified according to the classifier tool available at (http://www.evocell.org/hco/) [12].

Phylogenetic analyses

For phylogenetic analyses, multiple alignments were built with MAFFT v7.123b (option L-INS-i) [13] with the BLOSUM30, BLOSUM45 and BLOSUM62 matrix. The accuracy of the resulting alignments was verified with SEAVIEW 4.5.4 [14] and compared with NorMD v1.2 [15]. The best alignments were

trimmed with BMGE v1.1 [16] with the best BLOSUM matrix identified with NorMD and used for phylogenetic analyses. To avoid taxonomic redundancy, only one representative strain per bacterial genus was kept for phylogenetic reconstruction. The best fitting evolutionary models were identified by IQ-TREE (Nguyen et al. 2015) according to the BIC criterion [17]. Maximum Likelihood (ML) trees were built using PHYML v3.1 [18] with the NNI+SPR topology search option. The robustness of the resulting trees was estimated with a non-parametric bootstrap procedure (100 replicates of the original datasets).

The trees were visualized with iToL [19,20].

Analysis of genomic contexts

Genomic contexts were generated with GeneSpy developped by Pierre Garcia (personal communication).

Construction of a reference phylogeny of archaeal species

A reference phylogeny of *Archaea* was inferred using ribosomal proteins. The 70 ribosomal protein families available for 209 complete archaeal proteomes in RiboDB (https://ribodb.univ-lyon1.fr/ribodb/ribodb-in.cgi) [21] were retrieved and used as seed to query our local database. 39 on 391 available proteomes contained less than 35 ribosomal protein families. To limit biases due to missing data, the corresponding strains were not included in the phylogenetic analysis.

The 58 ribosomal proteins present in at least 70% of the 352 remaining proteomes were kept for phylogenetic analyses. For each ribosomal protein family a multiple alignment was built and trimmed as described above. The resulting datasets were combined together to build a large supermatrix containing 352 sequences, and 6,257 conserved amino acid positions. This supermatrix was used to inferred a ML tree with PHYML as described above.

Results

At least one HCO O_2-Red catalytic subunit was detected in most bacterial and archaeal lineages. A-HCO and its accessory subunits homologues were found in many *Archaea* and *Bacteria*. In contrast, B-HCO and its functional partners were mainly present in *Archaea*, while C-HCO and NOR were almost all present in *Bacteria*. For each subunit of the four O_2-Red families, we inferred the ancestral presence of the subunit in a taxon when the gene was widely distributed in this taxon and when we could find the monophyly of the group.

NOR catalytic subunit presents a scarce taxonomic distribution, being rarely present in more than 50% of proteomes of bacterial or archaeal order. NOR sequences from different taxa appeared intermixed on the tree, indicating that the relationships among NOR sequences were inconsistent with the currently recognized prokaryotic systematics. This combined to the scarce taxonomic distribution of NOR, indicated that their evolutionary history was heavily impacted by HGT, as proposed previously [2]. In fact, based on phylogenetic and taxonomic criteria, the ancestral presence of NOR can not be inferred in any lineage to the exception of *Halobacteria*. Thus, it was very unlikely that this enzyme could have been present in LUCA as proposed few years ago [3,7], or even in the ancestor of Bacteria or Archaea.

C-HCO catalytic subunit was widely distributed in the *Bacteroidetes*/*Chlorobi* group and in *Proteobacteria* that formed a well supported group, while they were scarce in other lineages. According to our data, the possible ancestral presence of C-HCO in the ancestor of *Chlorobi* and of *Proteobacteria*, two major bacterial lineages that didn't occupy a basal position in the phylogeny of *Bacteria* was not sufficient to infer its presence in the ancestor of Bacteria, as proposed recently [3]. The presence of these enzymes in strict anaerobic archaea was puzzling.

Most of the B-HCO homologues are found in *Archaea*, and more precisely in *Geoarchaeota*, *Sulfolobales* and *Halobacteria*. Regarding bacterial sequences, the phylogeny of B-HCO revealed a complex evolutionary history dominated by HGT between *Bacteria* and *Archaea* and among *Bacteria*. In particular, according to phylogenetic and taxonomic criteria it was not possible to trace back B-HCO in any bacterial lineage. In contrast, its ancestral presence in three major archaeal lineages suggested that it originated in this domain of Life, in agreement with previous studies [2,3].

According to taxonomic distribution and phylogenetic criteria, the ancestral presence of A-HCO could be inferred in *Cyanobacteria*, the *Bacteroidetes/Chlorobi* group, the *Fibrobacteres/Acidobacteria* group, *Alphaproteobacteria*, *Betaproteobacteria* and *Planctomycetes*. Other bacterial A-HCO appeared intermixed

suggesting that they spread through HGT in present-day bacterial lineages. Regarding Archaea, A-HCO homologues were widespread in several archaeal phyla, classes, and orders and formed monophyletic groups on the ML phylogeny. This was for instance the case of *Thaumarchaeota*, *Halobacteria*, *Geoarchaeota*, *Thermoproteales*, *Sulfolobales*, uncultured marine group II and Aigarchaeota, suggesting that they could be ancestral in these groups. The grouping of sequences from *Aigarchaeota*, *Geoarchaeota*, *Thermoproteales*, and *Sulfolobales*, close to a large *Thaumarchaeota* cluster, could indicate the ancestral presence of this enzyme in the ancestor of the recently proposed TACK superphylum. In contrast, the situation was less clear in the case of *Diaforarchaea*, because they formed a paraphyletic group that also included archaeal sequences from other taxa.

Beside catalytic subunits (subunit I), three additional subunits are shared by the A-HCO and B-HCO: the subunit II (SoxH, SoxA and DoxC), the cytochrome *b* (SoxG and SoxC) and the Rieske protein (SoxF and SoxL). The catalytic subunit and the subunit II co-occurred in almost all proteomes. Surprisingly, in 45% of *Halobacteria* strains, a copy of the subunit II was not co-localized with the catalytic subunit whereas an another copy did. At the exception of these Halobacteria, we can clearly distinguish the subunit II of the A-HCO and of the B-HCO in the phylogeny. We can highlight a topology similar to the one obtained with the catalytic subunit. Cytochrome b was not found in 30,7% of all proteomes of *Archaea* and complete genomes of *Bacteria* that contain a catalytic subunit. In the B-HCO, only the Sox of *Sulfolobales* had a copy of cytochrome b in the same genomic context as the catalytic subunit. The cytochrome b of A-HCO are in the same genomic context as the catalytic subunit only in *Crenarchaeota*. Rieske proteins were not found in 42,4% of all proteomes of Archaea and complete genomes of Bacteria that contain a catalytic subunit. The topology of Rieske protein is very similar to the topology of cytochrome b. It is not surprising because the two proteins are almost always in the same genomic context. Like the cytochrome b, Rieske protein is not in the same genomic context as the catalytic subunit at the exception of *Crenarchaeota*. The subunit III, subunit of A-HCO, was not found in 14,3% of all proteomes of *Archaea* and complete genomes of *Bacteria* that contain a catalytic subunit. The subunit III, in *Archaea*, is always fused with the catalytic subunit except for some *Halobacteria*. The subunit III present a similar topology to the A-HCO catalytic subunit at the exception that *Thaumarchaeota* do not have this subunit.

Regarding other subunits, the sulfocyanin (SoxE) is associated to A-HCO. The subunit is almost found in Archaea. More precisely, Sulfolobales sequences formed three monophyletic clusters, each of them could be traced back in the ancestor of this archaeal order. In contrast, the presence of SoxE in other archaeal lineages (e.g. *Thermoplasmatales*, *Aigarchaeota*, *Thermoproteales*) likely resulted from secondary HGT. We can highlight a curiosity about this subunit because for some strains, SoxE is in the same genomic context as the B-HCO catalytic subunit. SoxI was very likely present in the ancestors of *Geoarchaeota* and *Sulfolobales* and maybe also present in the ancestor of *Thermoproteales* and *Aigarchaeota*. This subunit is in the same genomic context as the A-HCO catalytic subunit. Interestingly, a similar cluster existed in the phylogeny of A-HCO, indicating that the presence of SoxI was co-opted specifically in the ancestor of this subgroup of A-HCO and preserved along its subsequent diversification. SoxD and DoxE associated to the Sox and Dox complexes (B-HCO), respectively, were found exclusively and in almost all *Sulfolobales*, suggesting they emerged in the stem of this order. Finally, the SoxD' subunit (B-HCO) and the Fox complex were restricted to a subgroup of *Sulfolobales* strains, indicating that they emerged late during the diversification of this archaeal order. Thus, they could not be traced back to the ancestor of this archaeal order and this implied that the corresponding proteins were much younger than the other subunits.

Discussion

Studying ancient metabolisms opens windows on primitive geosphere and biosphere, and a key to decipher the major transitions in the evolution of Earth [22]. Among them, the origin and rise of O_2 represent one of the most debated questions [23], specifically regarding the divergent conclusions drawn from biological and geological observations [23,24]. Present-day high level of O_2 in the atmosphere resulted of the oxygenic photosynthesis performed by *Cyanobacteria* [22,25–27] (including plant chloroplast that are of cyanobacterial origin), a process that started around 2.45-2.33 billion years ago [23,27,28]. Recent studies suggested that the tectonic activity around 2.7 billion years ago could have also contributed to release significant amounts of O_2 in the atmosphere (Lee et al. 2016; Sukumaran 2000; Jelen et al. 2016). The presence of small amount of O_2 of abiotic origin in primitive oceans and atmosphere prior 2.7 billion years remains debated [29–32]. In this context, studying the origin and evolution of HCO, the terminal electron acceptor of aerobic

respiratory chains, can provide precious clues for the emergence of this energetic process and the availability of O_2.

The survey of thousands of bacterial and archaeal proteomes encompassing all major lineages allowed inferring the presence of A-HCO in the ancestor of *Cyanobacteria*, *Bacteroidetes/Chlorobi*, *Fibrobacteres/Acidobacteria*, *Alphaproteobacteria*, *Betaproteobacteria*, *Planctomycetes*, and likely in Gammaproteobacteria. This confirmed previous results suggesting that A-HCO were ancient in *Bacteria* and may have predated the emergence of *Cyanobacteria* and thus of oxygenic photosynthesis [2]. In contrast and contrarily to previous proposals [7,33,34], we didn't find any evidence suggesting that B-HCO, C-HCO and NOR could be ancient in Bacteria.

Regarding *Archaea*, our data revealed a very dynamic evolution. First, we confirmed that A-HCO catalytic subunit and some of the accessory subunits were ancient in this Domain of Life [2,4]. More precisely, we showed that NOR could have been present in the ancestor of *Halobacteria*, B-HCO in the ancestors of *Sulfolobales* and *Halobacteria*, and A-HCO in the ancestors of *Thaumarchaeota*, *Crenarchaeota*, *Aigarchaeota*, uncultured marine group II/III and *Halobacteria*. Worth noticing, A-HCO were not restricted to mesophilic archaea, represented here by *Halobacteria*, uncultured marine group II/III and some *Thaumarchaeota*. This contradicted the recent proposal that the acquisition of the corresponding gene from bacterial donors could have been a crucial step towards the colonization of low-temperature environments by thermophilic *Archaea* [35]. On the contrary, A-HCO and B-HCO were widespread in thermophilic/hyperthermophilic archaeal lineages and we showed that B-HCO likely originated in *Archaea* and were subsequently transferred to a few *Bacteria*.

According to the rooting of the archaeal tree in-between the TACK superphylum and *Euryarchaeota*, it was proposed previously that A-HCO could be ancestral in *Archaea* [2]. Alternatively, one could propose that A-HCO and B-HCO were acquired independently and secondarily by HGT after the diversification of *Archaea* and the emergence of the main archaeal orders in response to the rise of O_2 level linked to the photosynthetic activity of *Cyanobacteria*. While, this conclusion was not contradict by our data, a new root splinting the Archaea into two clusters was recently proposed [36]. Cluster 1 corresponded to TACK, and two major euryarchaeotal lineages: the *Methanomada* (i.e. *Methanococcales*, *Methanobacteriales*, *Methanopyrales*) and the *Thermococcales*, while Cluster 2 encompassed *Diaforarchaea*, *Archaeoglobales*, ANME-1, *Halobacteria*, *Methanocellales*, *Methanosarcinales*, *Methanomicrobiales*. The position of this root opened the possibility that of a methanogenic and anaerobic ancestor for *Archaea* [36]. If confirmed, this would imply that the absence of any HGT of C-HCO from bacterial donor to archaea (excepted in two anaerobic methanosarcinales) was puzzling. However, independently of the root in *Archaea* our data indicate an ancient presence of A-HCO in *Archaea*.

In our study, we also carried out a thoughtful analysis of the other subunits included in the HCO complex in Archaea. Some subunits escorting the catalytic subunit, Cytochrome B, Rieske protein and the subunit II present the same evolutionary history than the catalytic subunit in Archaea and form the conserved core of A-HCO and B-HCO with the catalytic subunit. The subunit III is specific to the A-HCO complex and can be a part of the conserved core which has been lost by the ancestor of *Thaumarchaeota* and *Thermoplasmatales*. Interestingly, the other subunits have more recent origin and are a crenarchaeal innovation. Sulfocyanin (SoxE) of the A-HCO is particularly interesting because some copies are in the same genomic context as the B-HCO catalytic subunit. This result supports the hypothesis that the B-HCO could come from a duplication of the A-HCO in *Crenarchaeota* and explains that some specific subunits of A-HCO could work with B-HCO subunits. According to our results, it seems that the ancestor of TACK might have possessed two operons (A-HCO), one composed of the catalytic subunit, the subunit II and the subunit III and the other composed of the Rieske protein and the cytochrome b. During evolution, the subunit III was lost by the ancestor of *Thaumarchaeota* and *Thermoplasmatales* and the A-HCO complex has been duplicated in the ancestor of *Crenarchaeota* to give the B-HCO complex (with the lost of the subunit III) where the two operons get together. It is still unclear why Archaea and particularly Crenarchaeota innovate new subunits. Moreover, HCO was only studied in a few archaeal models and there is a lot of innovations in Archaea with accumulation of new subunits, so we are wondering if we have discovered all the proteins imply in the aerobic respiration in Archaea.

An other interesting point is the presence of O_2Red in strains or species described as strict anaerobes, such as *Methanosarcinales*, some *Deltaproteobacteria*, some *Gammaproteobacteria*, *Bacteroidetes*, some *Chlorobi*, *Chrysiogenetes* and *Deferribacteres*. This point has already highlight for other O2Red, the Cytochrome bd that was find also in species described as strict anaerobes (some *Deltaproteobacteria* and two

Methanosarcinales). It is likely that they are not pseudogenes. This suggest that these enzymes could be functional and could be, maybe, involved in detoxification [37].

Acknowledgements

We are grateful to Jose De La Torre (Department of Biology, San Francisco State University) for providing the sequences of HCO subunits of "*Canditatus* Nitrosocaldus yellowstonii". We thank Najwa Taïb for discussions and helpful comments on this work. We thank also Pierre Garcia for having supplied us his tool of visualization of genomic contexts (GeneSpy).

This work has been performed using the computing facilities of the CC LBBE/PRABI.

References

1. Pereira MM, Santana M, Teixeira M. A novel scenario for the evolution of haem-copper oxygen reductases. Biochim. Biophys. Acta [Internet]. 2001;1505:185–208. Available from: http://www.ncbi.nlm.nih.gov/pubmed/11334784

2. Brochier-Armanet C, Talla E, Gribaldo S. The multiple evolutionary histories of dioxygen reductases: Implications for the origin and evolution of aerobic respiration. Mol. Biol. Evol. 2009;26:285–97.

3. Ducluzeau A-L, Schoepp-Cothenet B, Lis R van, Baymann F, Russell MJ, Nitschke W. The evolution of respiratory O2/NO reductases: an out-of-the-phylogenetic-box perspective. J. R. Soc. Interface [Internet]. 2014 [cited 2014 Nov 7];11:20140196. Available from: http://rsif.royalsocietypublishing.org/content/11/98/20140196

4. Gribaldo S, Talla E, Brochier-Armanet C. Evolution of the haem copper oxidases superfamily: a rooting tale. Trends Biochem. Sci. [Internet]. 2009;34:375–81. Available from: http://www.ncbi.nlm.nih.gov/pubmed/19647436

5. Sousa FL, Alves RJ, Ribeiro MA, Pereira-Leal JB, Teixeira M, Pereira MM. The superfamily of heme-copper oxygen reductases: Types and evolutionary considerations. Biochim. Biophys. Acta - Bioenerg. 2012;1817:629–37.

6. Castresana J, Lübben M, Saraste M, Higgins DG. Evolution of cytochrome oxidase, an enzyme older than atmospheric oxygen. EMBO J. 1994;13:2516–25.

7. Ducluzeau A-L, van Lis R, Duval S, Schoepp-Cothenet B, Russell MJ, Nitschke W. Was nitric oxide the first deep electron sink? Trends Biochem. Sci. [Internet]. 2009;34:9–15. Available from: http://www.ncbi.nlm.nih.gov/pubmed/19008107

8. Rinke C, Schwientek P, Sczyrba A, Ivanova NN, Anderson IJ, Cheng J-F, et al. Insights into the phylogeny and coding potential of microbial dark matter. Nature [Internet]. Nature Research; 2013 [cited 2016 Dec 8];499:431–7. Available from: http://www.nature.com/doifinder/10.1038/nature12352

9. Brown CT, Hug LA, Thomas BC, Sharon I, Castelle CJ, Singh A, et al. Unusual biology across a group comprising more than 15% of domain Bacteria. Nature [Internet]. 2015;523:208–11. Available from: http://www.nature.com/doifinder/10.1038/nature14486

10. Altschul SF, Madden TL, Schäffer AA, Zhang J, Zhang Z, Miller W, et al. Gapped BLAST and PSI-BLAST: A new generation of protein database search programs. Nucleic Acids Res. [Internet]. 1997 [cited 2014 Jul 9];25:3389–402. Available from: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=146917&tool=pmcentrez&rendertype=abstract

11. Eddy SR. A new generation of homology search tools based on probabilistic inference. Genome Inform. [Internet]. 2009;23:205–11. Available from: http://www.ncbi.nlm.nih.gov/pubmed/20180275

12. Sousa FL, Alves RJ, Pereira-Leal JB, Teixeira M, Pereira MM. A Bioinformatics Classifier and Database for Heme-Copper Oxygen Reductases. PLoS One [Internet]. 2011 [cited 2014 Nov 5];6:e19117. Available from: http://dx.doi.org/10.1371/journal.pone.0019117

13. Katoh K, Misawa K, Kuma K, Miyata T. MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. Nucleic Acids Res. [Internet]. 2002;30:3059–66. Available from: http://www.ncbi.nlm.nih.gov/pubmed/12136088

14. Gouy M, Guindon S, Gascuel O. SeaView version 4: A multiplatform graphical user interface for sequence alignment and phylogenetic tree building. Mol. Biol. Evol. [Internet]. 2010;27:221–4. Available from: http://mbe.oxfordjournals.org/content/27/2/221.long

15. Thompson JD, Plewniak F, Ripp R, Thierry JC, Poch O. Towards a reliable objective function for multiple sequence alignments. J. Mol. Biol. [Internet]. 2001;314:937–51. Available from: http://www.sciencedirect.com/science/article/pii/S0022283601951873

16. Criscuolo A, Gribaldo S. BMGE (Block Mapping and Gathering with Entropy): a new software for selection of phylogenetic informative regions from multiple sequence alignments. BMC Evol. Biol. [Internet]. 2010 [cited 2015 Dec 7];10:1–21. Available from: http://link.springer.com/article/10.1186/1471-2148-10-210

17. Nguyen LT, Schmidt HA, Von Haeseler A, Minh BQ. IQ-TREE: A fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. Mol. Biol. Evol. 2015;32:268–74.

18. Guindon S, Dufayard J-F, Lefort V, Anisimova M. New Alogrithms and Methods to Estimate Maximum- Likelihoods Phylogenies: Assessing the performance of PhyML 3.0. Syst. Biol. 2010;59:307–21.

19. Letunic I, Bork P. Interactive Tree Of Life (iTOL): an online tool for phylogenetic tree display and annotation. Bioinformatics [Internet]. 2007 [cited 2016 Jan 2];23:127–8. Available from: http://www.ncbi.nlm.nih.gov/pubmed/17050570

20. Letunic I, Bork P. Interactive tree of life (iTOL) v3: an online tool for the display and annotation of phylogenetic and other trees. Nucleic Acids Res. [Internet]. Oxford University Press; 2016 [cited 2016 Sep 19];44:W242-5. Available from: http://www.ncbi.nlm.nih.gov/pubmed/27095192

21. Jauffrit F, Penel S, Delmotte S, Rey C, de Vienne DM, Gouy M, et al. RiboDB Database: A Comprehensive Resource for Prokaryotic Systematics. Mol. Biol. Evol. [Internet]. Oxford University Press; 2016 [cited 2016 Aug 30];33:2170–2. Available from: http://mbe.oxfordjournals.org/lookup/doi/10.1093/molbev/msw088

22. Jelen BI, Giovannelli D, Falkowski PG. The Role of Microbial Electron Transfer in the Coevolution of the Biosphere and Geosphere. Annu. Rev. Microbiol. [Internet]. Annual Reviews 4139 El Camino Way, PO Box 10139, Palo Alto, California 94303-0139, USA ; 2016 [cited 2016 Aug 19];70:annurev-micro-102215-095521. Available from: http://www.annualreviews.org/doi/10.1146/annurev-micro-102215-095521

23. Farquhar J, Zerkle AL, Bekker A. Geological constraints on the origin of oxygenic photosynthesis. Photosynth. Res. [Internet]. 2011 [cited 2017 Mar 17];107:11–36. Available from: http://www.ncbi.nlm.nih.gov/pubmed/20882345

24. Shih PM. Cyanobacterial Evolution: Fresh Insight into Ancient Questions. Curr. Biol. [Internet]. 2015 [cited 2017 May 9];25:R192–3. Available from: http://www.sciencedirect.com.inee.bib.cnrs.fr/science/article/pii/S0960982214016492

25. Lee C-TA, Yeung LY, McKenzie NR, Yokoyama Y, Ozaki K, Lenardic A. Two-step rise of atmospheric oxygen linked to the growth of continents. Nat. Geosci. [Internet]. 2016 [cited 2016 Aug 19];9:417–24. Available from: http://www.nature.com/doifinder/10.1038/ngeo2707

26. Sukumaran P V. Evolution of the atmosphere and oceans: Evidence from geological records. Resonance [Internet]. Springer India; 2000 [cited 2017 Jan 19];5:6–14. Available from: http://link.springer.com/10.1007/BF02834667

27. Lyons TW, Reinhard CT 1, Planavsky NJ 1. The rise of oxygen in Earth's early ocean and atmosphere. Nat. Febr. 20, 2014 [Internet]. 2014 [cited 2014 Nov 10];506:307–15. Available from: http://gate1.inist.fr/login?url=http://ovidsp.ovid.com/ovidweb.cgi?T=JS&CSC=Y&NEWS=N&PAGE=fulltext&D=ovfto&AN=00006056-201402200-00039

28. Luo G, Ono S, Beukes NJ, Wang DT, Xie S, Summons RE. Rapid oxygenation of Earths atmosphere 2.33 billion years ago. Sci. Adv. [Internet]. 2016 [cited 2017 Mar 13];2:e1600134–e1600134. Available from: http://advances.sciencemag.org/cgi/doi/10.1126/sciadv.1600134

29. Kendall B, Creaser RA, Reinhard CT, Lyons TW, Anbar AD. Transient episodes of mild environmental oxygenation and oxidative continental weathering during the late Archean. Sci. Adv. [Internet]. 2015;1:e1500777. Available from: http://advances.sciencemag.org/cgi/doi/10.1126/sciadv.1500777

30. Lu Z, Chang YC, Yin Q-Z, Ng CY, Jackson WM. Evidence for direct molecular oxygen production in CO2 photodissociation. Science (80-.). [Internet]. 2014 [cited 2014 Nov 5];346:61–4. Available from: http://www.ncbi.nlm.nih.gov/pubmed/25278605

31. Walker JCG. Oxygen and hydrogen in the primitive atmosphere. Pure Appl. Geophys. PAGEOPH [Internet]. Birkhäuser-Verlag; 1978 [cited 2017 Feb 28];116:222–31. Available from: http://link.springer.com/10.1007/BF01636879

32. Wang X-D, Gao X-F, Xuan C-J, Tian SX. Dissociative electron attachment to CO2 produces molecular oxygen. Nat. Chem. [Internet]. 2016 [cited 2017 Feb 28];8:258–63. Available from: http://www.ncbi.nlm.nih.gov/pubmed/26892558

33. Ducluzeau AL, Ouchane S, Nitschke W. The cbb3 oxidases are an ancient innovation of the domain bacteria. Mol. Biol. Evol. 2008;25:1158–66.

34. Ducluzeau A-L, Schoepp-Cothenet B, van Lis R, Baymann F, Russell MJ, Nitschke W. The evolution of respiratory O2/NO reductases: an out-of-the-phylogenetic-box perspective. J. R. Soc. Interface [Internet]. 2014;11:20140196. Available from: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=4233682&tool=pmcentrez&rendertype=abstract

35. López-García P, Zivanovic Y, Deschamps P, Moreira D. Bacterial gene import and mesophilic adaptation in archaea. Nat. Rev. Microbiol. [Internet]. 2015;13:447–56. Available from: http://www.ncbi.nlm.nih.gov/pubmed/26075362

36. Raymann K, Brochier-Armanet C, Gribaldo S. The two-domain tree of life is linked to a new root for the Archaea. Proc. Natl. Acad. Sci. U. S. A. [Internet]. National Academy of Sciences; 2015 [cited 2016 Aug 19];112:6670–5. Available from: http://www.ncbi.nlm.nih.gov/pubmed/25964353

37. Le Fourn C, Brasseur G, Brochier-Armanet C, Pieulle L, Brioukhanov A, Ollivier B, et al. An oxygen reduction chain in the hyperthermophilic anaerobe Thermotoga maritima highlights horizontal gene transfer between Thermococcales and Thermotogales. Environ. Microbiol. [Internet]. 2011 [cited 2017 Feb 27];13:2132–45. Available from: http://doi.wiley.com/10.1111/j.1462-2920.2011.02439.x

Genomicus – New tools for comparative genomics and evolution in eukaryotes

Alexandra LOUIS[1], Nga Thi Thuy NGUYEN[1], Hugues ROEST CROLLIUS[1]

[1] Ecole Normale Supérieure, Inserm, CNRS, IBENS, F-75005 Paris, France

Corresponding Author: alouis@biologie.ens.fr

Since 2010, the Genomicus [1,2] web server is available online at http://genomicus.biologie.ens.fr. This graphical browser provides access to comparative genomics analyses in four different phyla (Vertebrate, Plants, Fungi, and non vertebrate Metazoans). Users can manipulate evolutionary genomic information from extant species, as well as ancestral gene content and gene order for vertebrates and flowering plants.

Annotated genes are presented in a classical but highly customisable phylogenetic framework, while harnessing the added signal provided by local gene organisation. Gene presence/absence, Ka and Ks, % Id, genomic distance are available in addition to homology relationships. Genomicus is used by a broad community of users that perform >30,000 individual queries per months.

New analysis and visualisation tools have recently been implemented in Genomicus Vertebrate. Entire genomes (karyotype structures) can now be compared between multiple genomes, and synteny blocks can now be computed and visualised between any two genomes [3].

In this tutorial we will take users through the 6 different visualisation and analysis tools via a number of real test cases, aimed at answering questions on karyotype evolution, gene synteny conservation, gene duplication, gene gain, gene loss and chromosomal rearrangement.

References

[1] Louis, A., Murat, F., Salse, J. and Roest Crollius, H. GenomicusPlants: A Web Resource to Study Genome Evolution in Flowering Plants. *Plant Cell Physiol*. 56: e4. 2015

[2] Louis, A., Nguyen, N.T., Muffato, M. and Roest Crollius, H. Genomicus update 2015: KaryoView and MatrixView provide a genome-wide perspective to multispecies comparative genomics. *Nucleic acids research* 43: D682-689. 2015

[3] Lucas, J. M., Muffato, M. & Roest Crollius, H. PhylDiag: identifying complex synteny blocks that include tandem duplications using phylogenetic gene trees. *BMC Bioinformatics* 15, 268 (2014).

Conférence invitée

Céline Brochier

Université Lyon 1, France

The growing tree of Archaea: changing perspectives on the diversity and evolution of the third domain of life

Archaea occupy a key position in the Tree of Life, and represent a major fraction of the microbial diversity. Abundant in soils, ocean sediments and the water column, they are key players in processes mediating global carbon and nutrient fluxes, as well as important components of the animal microbiome and human body. The development of culture-independent sequencing techniques has revealed a myriad of so far inaccessible microbial lineages and filled up the archaeal tree with entirely new branches. The unprecedented access to genomic data from a large number of archaeal lineages provides the raw material for dissecting the origin of this domain, the evolutionary trajectories that have shaped its current diversity, and its relationships with Bacteria and Eukaryotes. This rainfall of data combined to cutting-edge methods allowing to disentangle the multiple signals contained in molecular sequences has shed new light on the evolutionary history of Archaea. Here I will review the major advances in the field as well as important open issues and future challenges.

Probing factor-dependent long-range contacts using regression with higher-order interaction terms

Raphaël MOURAD[1], Lang LI[2] and Olivier CUVIER[1]
[1] Laboratoire de Biologie Moléculaire Eucaryote (LBME), CNRS, Université Paul Sabatier (UPS), 31000 Toulouse, France
[2] Center for Computational Biology and Bioinformatics (CCBB), Indiana University, 46202 Indianapolis, USA

Corresponding author: raphael.mourad@ibcg.biotoul.fr

Reference paper: **Mourad *et al.* (2017). Uncovering direct and indirect molecular determinants of chromatin loops using a computational integrative approach. PLOS Computational Biology, *in press*.**

Abstract *Chromosomal organization in 3D plays a central role in regulating cell-type specific transcriptional and DNA replication timing programs. Yet it remains unclear to what extent the resulting long-range contacts depend on specific molecular drivers. Here we propose a model that comprehensively assesses the influence on contacts of DNA-binding proteins, cis-regulatory elements and DNA consensus motifs. Using real data, we validate a large number of predictions for long-range contacts involving known architectural proteins and DNA motifs. Our model outperforms existing approaches including enrichment test, random forests and correlation, and it uncovers numerous novel long-range contacts in Drosophila and human.*

Keywords Epigenetics; Chromatin; Hi-C; ChIP-seq; Generalized linear model

1 Introduction

The comprehensive analysis of 3D chromatin drivers is currently a hot topic [1]. A growing body of evidence supports the role of insulator binding proteins (IBPs) such as CTCF, and cofactors like cohesin, as mediators of long-range chromatin contacts [2]. In this work, we used a generalized linear model with interactions (GLMI) to identify the molecular determinants of loops, including protein and DNA sequence. Using this model, we uncovered numerous novel DNA loops and underlying mechanisms in *Drosophila* and human.

2 Results

The model is formulated as follows:

$$\begin{aligned}\log\left(\mathrm{E}\left[\mathbf{y}|\mathbf{X}\right]\right) &= \beta_0 + \boldsymbol{\beta}\mathbf{X} \\ &= \beta_0 + \beta_d\mathbf{d} + \boldsymbol{\beta}_B\mathbf{B} + \boldsymbol{\beta}_C\mathbf{C} + \beta_g\mathbf{g} \end{aligned} \quad (1)$$

Variable $\mathbf{y}$ denotes the number of high-throughput chromosome conformation capture (Hi-C) contacts for any pair of bins on the same chromosome. Variable set $\mathbf{X} = \{\mathbf{d}, \mathbf{B}, \mathbf{C}, \mathbf{g}\}$ comprises several variable subsets: the log-distance variable $\mathbf{d}$ (polymer effect), the bias variables $\mathbf{B}$ (GC content, fragment length and mappability), the confounding variable set $\mathbf{C}$ and the genomic variable of interest $\mathbf{g}$. Model (1) is very general and can be developed in multiple versions depending on the variable $\mathbf{g}$ of interest. In this highlight article, we will only present the simplest version of the model, although more sophisticated versions accounting for multiple proteins mediating loops have been developed. The model is available in the R package "HiCglmi" which can be downloaded from the Comprehensive R Archive Network.

Let consider a pair of bins that we call left bin (L) and right bin (R). The attribution for left and right bins is arbitrary. Let also consider a genomic feature F_i (whose binding is colored in blue in Figure 1a), that represent binding sites of the same protein. For the genomic feature F_i, occupancy variables $\mathbf{z}_{iL}$ and $\mathbf{z}_{iR}$ denote the occupancies of F_i on left and right bins, respectively. For an occupancy variable, a value of 0/1 means absence/presence of the corresponding feature on the bin, *e.g.* absence/presence of the protein on the bin (a value between 0 and 1 means partial overlap of the feature). A "homologous interaction" variable $\mathbf{g} = \mathbf{n}_{ii}$ is the product of $\mathbf{z}_{iL}$ and $\mathbf{z}_{iR}$ ($\mathbf{n}_{ii} = \mathbf{z}_{iL} \times \mathbf{z}_{iR}$). The associated $\beta_{n_{ii}}$ parameter reflects the extent by which the genomic feature F_i interacts with itself through chromatin contacts (Figure 1a). For instance, distant CTCF

binding sites were shown to form loops in human [2].

We compared GLMI with existing methods for their ability to identify genomic features known to be involved in long-range contacts (Figure 1b). Here we used the negative binomial regression as the best specification of the GLMI in the context of Hi-C data overdispersion. We compared GLMI with (1) enrichment test (ET) on highly confident chromatin interaction pairs as previously [3], (2) correlation (Cor) on highly confident chromatin interaction pairs [4] and (3) random forests (RF) discriminating highly confident chromatin interaction pairs from non-interacting pairs [5]. We found that GLMI outperformed the other methods to detect long-range contacts between known architectural protein binding motifs. Using GLMI, we also uncovered novel long-range contacts between architectural proteins in *Drosophila* (Figure 1c) and in human (data not shown).

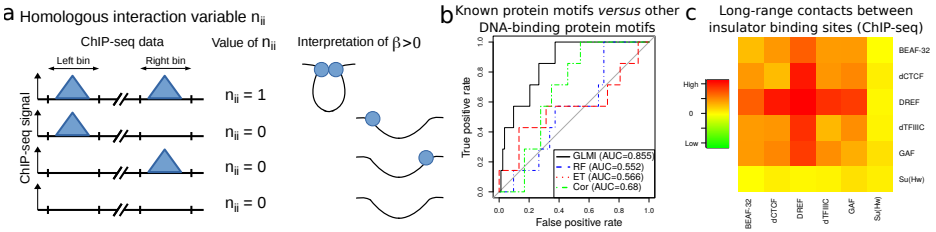


Fig. 1. a) Illustration of homologous interaction variable. b) Comparison between GLMI, random forest (RF), enrichment test (ET) and correlation (Cor) to detect known long-range contacts between protein motifs. c) Heatmap of long-range contacts between architectural proteins in *Drosophila*.

3 Conclusion

Here, we propose to use a generalized linear regression with interactions (GLMI) to study the roles of genomic features such as DNA-binding proteins, motifs or promoters to bridge long-range contacts in the genome, depending on transcriptional status or motif orientation. GLMI has multiple assets over existing approaches such as enrichment test, correlation and random forests. Compared to enrichment test or correlation that respectively assesses the protein enrichment or correlation at highly confident loops, GLMI quantitatively links the frequency of all long-range contacts to complex co-occupancies of proteins while accounting for known Hi-C biases and polymer background. Moreover, GLMI accounts for colocalizations among protein binding, a strong issue when analyzing protein binding sites known to largely overlap over the genome. In contrast to random forests which are efficient predictive models but sometimes poor explanatory ones, GLMI allows to identify key chromatin loop driver proteins and motifs. GLMI can also uncover numerous mechanisms behind loop formation using higher-order interaction terms and proper confounding variables (see original article). For instance, GLMI can determine if a cofactor is necessary to mediate long-range contacts between distant protein binding sites.

References

[1] Caelin Cubenas-Potts and Victor G. Corces. Architectural proteins, transcription, and the three-dimensional organization of the genome. *FEBS Letters*, 589(20PartA):2923–2930, 2015.

[2] Suhas S. P. Rao, Miriam H. Huntley, Neva C. Durand, Elena K. Stamenova, Ivan D. Bochkov, James T. Robinson, Adrian L. Sanborn, Ido Machol, Arina D. Omer, Eric S. Lander, and Erez L. Aiden. A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping. *Cell*, 159(7):1665–1680, February 2015.

[3] Ka-Chun Wong, Yue Li, and Chengbin Peng. Identification of coupling dna motif pairs on long-range chromatin interactions in human k562 cells. *Bioinformatics*, 2015.

[4] Vera Pancaldi, Enrique Carrillo-de Santa-Pau, Biola Maria Javierre, David Juan, Peter Fraser, Mikhail Spivakov, Alfonso Valencia, and Daniel Rico. Integrating epigenomic data and 3D genomic structure with a new measure of chromatin assortativity. *Genome Biology*, 17(1):1–19, 2016.

[5] Bing He, Changya Chen, Li Teng, and Kai Tan. Global view of enhancer-promoter interactome in human cells. *Proceedings of the National Academy of Sciences of the United States of America*, 111(21):201320308–E2199, May 2014.

An integrative approach for predicting the RNA secondary structure for the HIV–1 Gag UTR using probing data

Afaf SAAIDI[1] supervised by Yann PONTY[1] and Bruno SARGUEIL[2]
[1] AMIBio team, Laboratoire d'informatique de l'école polytechnique, Inria Saclay, 91120, Palaiseau, France
[2] Laboratoire de cristallographie et RMN Biologiques, Faculté de pharmacie Paris Descartes, 75006, Paris, France

Corresponding author: yann.ponty@lix.polytechnique.fr

Structure modeling is key to understand the mechanisms of RNA retroviruses such as HIV. Many *in silico* prediction approaches suggesting structural models of moderate to good accuracies are available. However, the prediction methods could be further improved by taking advantage of both next generation sequencing technologies and different experimental techniques such as enzymatic and SHAPE probing data [1]. In a published article [2], we introduce and use a structural modeling method based on the integration of many experimental probing data to direct predictions with the aim to find the most accurate structure lying in the intersection of disjoint sources of experiments.

Method. High-throughput experimental data can be derived from two sources of experimental data: SHAPE [1] and enzymatic probing. We used the stochastic sampling [3] mode of `RNASubopt` to sample structural models from the Boltzmann distribution in a way that favors/forces compatibility with the derived constraints. Namely, SHAPE reactivity profiles were used as soft constraints, meaning that observed reactivity values were translated into pseudo-energies. Position-specific susceptibilities to RNAses cleavage were used as hard constraints by setting arbitrary cut-offs above which specific base are forced to be paired and unpaired. Both types of constraints reduce the space of possible conformations, leading to a set of structures that are maximally compatible with the provided data.

We posited that the optimal structure(s) should be energetically stable and supported by several experimental data. Thus, for each type of probe, we generate a set of structures compatible with experimentally-derived constraints. We merge those sets, and performed a structural distance-based clustering across experimental conditions, to generate several sets of structural models that are well-supported by experimental data. Clusters were scored using three criteria, namely, their stability, coherence and diversity (recurrence across structural conditions). Within the set of clusters returned by our clustering algorithm we elect clusters on the Pareto frontier, ie clusters that are not strictly dominated with respect to these three criteria. Finally, representative structures (centroids), corresponding to Maximum Expected Accuracy structures are built and returned.

Results. The HIV-1 sequence probed in this study corresponds to the 5'UTR preceding the gag coding region from the NL-4.3 strain (Genbank: AF324493.2). A sample set of 12 000 structures, covering 6 sources of experimental data, was generated. The clustering step led us to elect two optimal clusters, whose corresponding centroids were then assessed in the light of their compatibility with specific SHAPE data. This allowed us narrow down our proposed models to a single candidate, whose base pair conservation/covariation was confirmed by comparative analysis.

Conclusion and perspectives. Our integrative approach allowed us to implement a consensus structure compatible with many different experimental probing data. As further work, we project to use our approach with other sources of experimental probing data and to exploit the alignment to build an additional constrained sample instead of using it for validation.

References

[1] K. A Wilkinson, E. J Merino, and K. M Weeks. Selective 2'-hydroxyl acylation analyzed by primer extension (SHAPE): quantitative RNA structure analysis at single nucleotide resolution. *Nature protocols*, 1(3):1610–6, 2006.

[2] J. Deforges, S. de Breyne, M. Ameur, N. Ulryck, N. Chamond, A. Saaidi, Y. Ponty, T. Ohlmann, and B. Sargueil. Two ribosome recruitment sites direct multiple translation events within HIV1 Gag open reading frame. *Nucleic Acids Research*, 2017.

[3] Y. Ding and C Lawrence. A statistical sampling algorithm for RNA secondary structure prediction. *Nucleic Acids Research*, 31(24):7280–7301, 2003.

NCboost: a meta-classifier of pathogenic non-coding variants integrating multiple sequence conservation and unsupervised functional scores

Barthelemy CARON[1], Lotfi SLIM[1] and Antonio RAUSELL[1]

[1] Clinical BioInformatics Lab - Imagine Institute, 24 Boulevard du Montparnasse, 75015, Paris, France

Corresponding Author: barthelemy.caron@institutimagine.org

Whole Genome Sequencing (WGS) allows the identification of rare variants in non-coding regions with potential regulatory consequences. In the study of rare genetic diseases, where the number of sequenced individuals is typically low, computational approaches are needed to prioritize variants for further functional follow-up. Over the last years, a number of methods have been developed to predict pathogenicity of non-coding variants (e.g. CADD, GWAVA, FunSeq2[1]). Such methods are often based on machine-learning algorithms trained on curated disease-causing variants, and exploit features such as DNA methylation, histone modifications and transcription factor binding sites . However, their predictive performance is still low and has been shown to mainly rely on sequence conservation scores.

Here, we present NCBoost, an XGBoost-based [2] meta-classifier, trained on non-coding pathogenic variants from the Human Gene Mutation Database (HGMD). NCboost exploits a comprehensive set of purifying selection signals at three levels: the affected position, the surrounding non-coding region and the closest gene. In order to propose an integrative strategy for non-coding variants prioritization, we additionally included the functional scores of two unsupervised methods for the detection of regulatory variants: DeepSEA[3] and Eigen[4], both of them integrating epigenetic functional and conservation features.

NCboost performance was evaluated both by 10-fold cross validation and against an independent set of pathogenic variants from ClinVar database. A detailed comparative benchmark against state-of-the art methods is presented and results discussed in terms of the quality of the annotation sets and the targeted regulatory regions.

Acknowledgements

We thanks YuFei Luo for her contribution in variant annotation.

References

[1] YaoFu et al, FunSeq2: a framework for prioritizing noncoding regulatory variants in cancer, Genome Biology, 2014.
[2] Chen,T. and Guestrin,C. (2016) XGBoost: A Scalable Tree Boosting System. ACM Press, pp. 785–794.
[3] Zhou,J. and Troyanskaya,O.G., Predicting effects of noncoding variants with deep learning-based sequence model. Nat. Methods , 12 , 931–934, 2015.
[4] Iuliana Ionita-Laza et al, A spectral approach integrating functional genomic annotations for coding and non coding variants, Nat.Gen. 2016

Context-specific prioritization of non-coding variants implicated in human diseases

Lambert MOYON[1], Yves CLÉMENT[1] , Camille BERTHELOT[1] and Hugues ROEST CROLLIUS[1]

[1] DYOGEN TEAM, IBENS, CNRS UMR8197, 46 rue d'Ulm, 75005, Paris, France

Corresponding Author: hugues.roest.crollius@ens.fr

High-quality whole genome sequences (WGS) of patients with rare or common diseases are increasingly being used to search for causative genetic variants that may explain the disease phenotype. In a large fraction of cases, no coding mutation in a known disease gene can be found, raising the possibility of at least two alternative causes: (1) a deleterious coding mutation in a new disease gene (2) a non-coding variant, for example modifying the function of an enhancer or a promoter of a disease gene. In the first case, a candidate-gene approach can be employed to investigate additional possibilities. In the second case, while methods exist to automatically annotate variants as potentially overlapping a regulatory region, investigators remain at loss for a reliable guide to efficiently prioritize the many variants that generally fall in this category, especially with respect to their impact on the patient's disease phenotype.

We present here a method aiming at prioritizing non-coding variants in a disease-oriented manner. By integrating functional, biochemical, and evolutionary information, our goal is to identify variants that are likely to be functional, with regards to the considered phenotype. This phenotype-oriented prioritization relies on the use of resources that link regulatory regions with target genes. In the context of a disease with known curated genes, this enables us to yield predictions of potentially causal variants that can be further investigated.

To evaluate our method, we are working in collaboration with the BRIDGE project from the University of Cambridge. More than 9,000 high-quality whole genome sequenceus from patients affected by rare diseases are now available, with a wide range of phenotypes, organized in 15 sub-projects. We applied our classification pipeline to cohorts of patients grouped by disease-phenotypes, and present here the specific properties of the predicted causal variants.

References

[1] Albert, F.W., and Kruglyak, L. (2015). The role of regulatory variation in complex traits and disease. Nat Rev Genet 16, 197–212.

[2] Carss, K.J., Arno, G., Erwood, M., Stephens, J., Sanchis-Juan, A., Hull, S., Megy, K., Grozeva, D., Dewhurst, E., Malka, S., et al. (2017). Comprehensive Rare Variant Analysis via Whole-Genome Sequencing to Determine the Molecular Pathology of Inherited Retinal Disease. The American Journal of Human Genetics 100, 75–90.

Posters

Analysis workflow for smallRNA-seq data

Sébastien NIN[1], Stéphanie RIALLE[1], Emeric DUBOIS[1] AND Laurent JOURNOT[1]

[1] Plateforme MGX - Montpellier GenomiX, c/o Institut de Génomique Fonctionnelle, 141, rue de la cardonille, 34094 MONTPELLIER Cedex 05, FRANCE

Corresponding author: emeric.dubois@mgx.cnrs.fr

Abstract

S*mall RNAs* are small sequences, generally non-coding and playing a role in gene silencing effects, thus allowing the inhibition of the regulation or the translation of their target messenger RNA.

After the sequencing, reads obtained from small RNA libraries require a specific bioinformatics treatment in order to identify and quantify small RNAs. Here we focus on the analysis of micro RNAs.

After a quality control step, the 3' adapter sequence is trimmed from the reads (Cutadapt). Then, miRDeep2[1], a software able to quantify, identify micro RNAs (miRNAs) and discover new miRNAs through a secondary structure analysis, is used. It first maps the reads to a reference genome allowing 5 multi-mapping positions, then quantifies the miRNAs. We compared miRDeep2 to other tools and chose it for its performance in term of quantification of known miRNAs. From the quantification results, a gene differential expression analysis (edgeR[2], DESeq[3], DESeq2[4]) is done. The analysis ends with a target mRNA prediction (miRGate) and functionnal enrichment of Gene Ontology terms with TopGO (Bioconductor).

The workflow is used on the Montpellier GenomiX facility to provide a new analysis service.

Citations

[1] Friedländer MR et al., miRDeep2 accurately identifies known and hundreds of novel microRNA genes in seven animal clades, Nucleic acids research 2012, Vol. 40, No 1: 37-52

[2] Robinson MD, McCarthy DJ, Smyth GK. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics*. 2010;26(1):139-140. doi:10.1093/bioinformatics/btp616

[3] Anders et al., Differential expression analysis for sequence count data, Genome Biology 2010, 11:R106 DOI: 10.1186/gb-2010-11-10-r106

[4] Love et al., 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

Analyse du réseau moléculaire impliqué dans le remodelage ventriculaire gauche post-infarctus du myocarde

Marie CUVELLIEZ[1], Christophe BAUTERS[2], Thomas KELDER[3], Marijana RADONJIC[3],Philippe AMOUYEL[1]and Florence PINET[1]

[1]Inserm U1167, Institut Pasteur de Lille, 1 rue du Pr Calmette, 59019 Lille, France
[2]Inserm U1167, CHRU Lille, 2 avenue Oscar Lambret, 59037 Lille, France
[3]EdgeLeap B.V.,Hooghiemstraplein 15, 3514 AX Utrecht, Pays Pas

Corresponding Author: florence.pinet@pasteur-lille.fr

Les maladies cardiovasculaires sont l'une des principales causes de mortalité dans les pays occidentaux. Suite à un infarctus du myocarde (IDM), 30% des patients développent un remodelage ventriculaire gauche (RVG) pouvant aboutir à une insuffisance cardiaque (IC).

REVE-2 est une étude multicentrique de 246 patients hospitalisés pour un premier IDM antérieur. Un suivi échocardiographique et des prélèvements sanguins sont réalisés à 4 temps après l'IDM : à l'inclusion (baseline), 1 mois, 3 mois et 1 an. A chaque temps, 25 variables moléculaires ont été dosées dans le plasma des patients (19 protéines et 6 ARNs non codants). L'expression de ces variables a été comparée entre les patients avec un RVG élevé (>20%) et un RVG faible (<20%) calculé entre l'échographie à un an et à l'inclusion.

Un réseau basé sur les interactions moléculaires connues a été construit à partir de l'ensemble des variables moléculaires dosées dans le plasma des patients de REVE-2. Pour cela, la plateforme EdgeBox (EdgeLeap), composée de 12 bases de données publiques (ENCODE, EnsemblGenes, HMDB, Microcosm, miRBase, miRecords, miRTarBase, Reactome, STRING, TargetScan, Tfe et WikiPathways) a été utilisée. Le réseau inclut les variables REVE-2, les voisins directs de ces molécules et les molécules faisant partie des chemins les plus courts lorsque 3 interactions relient 2 variables REVE-2. Le réseau appelé REVE-2 est constitué de 1310 molécules dont 1263 protéines, 24 microARNs, 22 métabolites et un ARN long non codant. L'analyse du réseau a identifié 40 clusters qui ont été annotés pour les processus biologiques dans Gene Ontology (GO). Le réseau est visualisé sous Cytoscape (version 3.2.1). Une analyse des modules actifs a été réalisée pour chaque temps (inclusion, 1 mois, 3 mois et 1 an) à partir des variables REVE-2 significativement modulées entre les 2 groupes de patients (Pinet et al., 2017). La majorité des changements d'expression des molécules du réseau est observée en baseline et à 3 mois post-IDM, correspondant respectivement à la phase post-IDM et au développement du RVG. Une analyse de la centralité de chaque molécule a permis de déterminer leur importance dans le réseau, une centralité élevée suggérant un rôle crucial de la molécule dans le processus physiopathologique. Cette analyse a permis d'identifier de nouvelles molécules, non encore quantifiées chez les patients REVE-2, potentiellement impliquées dans le développement du RVG post-IDM. Les facteurs de transcription EP300, CTCF et ESR1 sont exprimés dans le cœur et ont été identifiés comme régulant l'activité de SOD2, protéine intervenant dans la régulation du stress oxydant dans les cardiomyocytes. Quatre ARNs non codants, les miR-26b-5p, miR-17-5p, miR-335-5p et miR-375, ont une centralité importante. Les deux premiers sont exprimés dans le cœur, contrairement aux 2 derniers. Les taux plasmatiques du miR-26b-5p sont diminués chez les patients avec une IC aigue et le miR-17-5p est impliqué dans l'apoptose des cardiomyocytes et la fibrose cardiaque.

L'analyse du réseau REVE-2 permet d'étudier les mécanismes physiopathologiques associés au RVG post-IDM au cours du temps, ainsi que d'identifier de nouveaux biomarqueurs potentiels pour détecter le RVG associé à un risque élevé d'IC.

Mitochondrial genome variability of 205 Arabian endurance horses

Alexandre HEURTEAU[1], Claire HOEDE[1], Anne RICARD[2, 4], Diane ESQUERRÉ[3], Caroline MORGENTHALER[4], Nuria MACH[4], Céline ROBERT[4, 5], Eric BARREY[4],

[1] PF Bioinfo Genotoul, MIAT, Université de Toulouse, INRA, Castanet-Tolosan, France
[2] IFCE, Recherche et innovation, 61310 Exmes
[3] GeT PlaGe, Genotoul, INRA Auzeville, Castanet Tolosan, France
[4] GABI, BIGE, INRA, AgroParisTech, Université Paris-Saclay, Jouy-en-Josas, France
[5] Ecole Nationale Vétérinaire d'Alfort, Université Paris-Est, Maison-Alfort, France

Corresponding Author: eric.barrey@inra.fr

1 Introduction

Endurance horses can run up to 160 km per race. Mitochondrial DNA variations may affect the efficiency of electron transport chain and ROS (Reactive Oxygen Species ie: peroxide) production, thus contributing to endurance performance. We studied the mitochondrial genome variability among 205 endurance horses for which racing performance were known. The objectives of this study were to propose a new strategy to call genetic variants in mitochondrial genome, and determine whether some variants are associated to endurance performance.

2 Material & methods

Sampling & sequencing: We designed 5 overlapping amplicons to specifically amplify mitochondrial genome from total DNA extracted in peripheral blood. Then we sequenced all samples with Illumina Miseq that produces 250 bp paired reads.

Bioinformatics: We chose a Arabian reference genome (GenBank ID JN398380.1) that we pseudo-circularized in silico to improve terminals' alignments. The cleaning and mapping pipeline is standard. We filtered reads with mapping quality < 60. We used the GATK package [1] to detect variations. Most parameters followed the best practices and some were fine-tuned.

Statistics: Association between SNPs and performance in endurance was tested using mixed model with fixed SNP Effect and random additive genetic effect with relationship matrix.

3 Results & conclusion

For the 205 horses, we found 590 combined variable positions (one position varies every 27 bps in average). Only 5 contain indels. 72% have already been found in previous studies. 80% of the protein coding variants are silent and the transition/transversion ratio is 22.5. Furthermore, we could observe 1.5% of non-haploid genotype (potential heteroplasmy) but some of them are probably numts (nuclear sequences of mitochondrial origin [2]). Preliminary study of statistical analysis can not yet link any variants to performance in endurance racing. Further analysis is still required.

References

[1] Aaron McKenna, Matthew Hanna, Eric Banks, Andrey Sivachenko, Kristian Cibulskis, Andrew Kernytsky, Kiran Garimella, David Altshuler, Stacey Gabriel, Mark Daly, Mark A DePristo. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. Genome Research, 20:1297-303, 2010.

[2] Solomon G. Nergadze, Manuel Lupotto, P. Pellanda, Marco Santagostino, Valerio Vitelli and Elena Giulotto. Mitochondrial DNA insertions in the nuclear horse genome. Animal Genetics, 41: 176–185, 2010.

Epigenetic heterogeneity in multiple myeloma

Jennifer RONDINEAU[1], Victor GABORIT[1,2], Catherine GUERIN-CHARBONNEL[1,3], Philippe MOREAU[1,4], Stéphane MINVIELLE[1,4], and Florence MAGRANGEAS[1,4]

[1] CRCINA, INSERM, CNRS, Université d'Angers, Université de Nantes, Nantes, France
[2] LS2N, CNRS, Université de Nantes, Nantes, France
[3] Institut de Cancérologie de l'Ouest, Nantes, France
[4] CHU de Nantes, Nantes, France

Corresponding Author: jennifer.rondineau@univ-nantes.fr

Multiple myeloma (MM) is a hematological malignancy characterized by proliferation of malignant plasma cells. Despite the evolution of treatments over the last few years, MM remains an incurable disease, the vast majority of patients relapse of their disease in the years following treatment. MM represents about 20% of deaths from hematologic malignancies and 2% of cancer deaths. Although many types of genetic lesions (translocation, deletions, amplifications, mutations) have been identified, they don't fully explain the molecular mechanisms of relapse. We hypothesized that epigenetic changes contribute to the resistance of MM. Modified epigenetic genes could drive a novel mechanism of drug resistance in MM through changes in epigenetic states, subclonality and diversity. To test this hypothesis, we sought to highlight changes in the composition and diversity of epigenetic alleles during the progression of MM.

We used ERRBS (Enhanced Reduced Representation Bisulfite Sequencing) technique to analyze the DNA methylome of 17 patients at diagnosis and their relapse of MM, and 3 normal plasma cells (NPC) as control. This technique covers about 6% of CpGs of the genome and provides a single base resolution. Bisulfite reads were aligned using Bismark aligment software [1] to the bisulfite-converted hg19 genome, with non-directional model. We calculated the epigenetic changes between two stages using methclone [2]. This tool detects locus of 4 adjacent CpGs (minimum depth of 60X), called epiallele. We were particularly interested in the epiallele shift at relapse compared to diagnosis. Epigenetically shifted loci (eloci) are defined by a significant entropy shift ($\Delta S < -70$).

We showed that epigenetic allelic diversification occurs during the initiation of the disease and also during progression and is highly variable between patients. Moreover, localization of eloci indicate important perturbations of regions with potential regulatory effects on gene expression during MM progression. These results are similar to observations made in other tumors, such as acute myeloid leukemia [3].

Acknowledgments

This work is supported by the "Fondation Française pour la Recherche contre le Myélome et les Gammapathies monoclonales".

References

[1] KRUEGER, Felix et ANDREWS, Simon R. Bismark: a flexible aligner and methylation caller for Bisulfite-Seq applications. *bioinformatics*, 2011, vol. 27, no 11, p. 1571-1572.

[2] LI, Sheng, GARRETT-BAKELMAN, Francine, PERL, Alexander E., *et al.* Dynamic evolution of clonal epialleles revealed by methclone. *Genome biology*, 2014, vol. 15, no 9, p. 472

[3] LI, Sheng, GARRETT-BAKELMAN, Francine E., CHUNG, Stephen S., *et al.* Distinct evolution and dynamics of epigenetic and genetic heterogeneity in acute myeloid leukemia. *Nature Medicine*, 2016, vol. 22, no 7, p. 792-799.

Keywords

Multiple Myeloma, epigenetic, whole genome DNA methylation, tumor heterogeneity

NF-κB Landscape in Multiple Myeloma by high-throughput sequencing analysis

Victor GABORIT[1,2], Jennifer RONDINEAU[1], Wilfried GOURAUD[1,3], Jérémie BOURDON[2], Stéphane MINVIELLE[1,4] and Florence MAGRANGEAS[1,4]

[1] CRCINA, INSERM, CNRS, Université d'Angers, Université de Nantes, Nantes, France
[2] LS2N, CNRS, Université de Nantes, Nantes, France
[3] Institut de Cancérologie de l'Ouest, Nantes, France
[4] CHU de Nantes, Nantes, France

Corresponding author: victor.gaborit@univ-nantes.fr

Nuclear factor κB (NF-κB) subunits RelA, RelB, cRel, p50 and p52 are each critical for B cell function such as regulation of cell proliferation or immune response. The NF-κB pathway is constitutively activated for more than 20% of patients with multiple myeloma (MM), a hematological malignancy characterized by an accumulation of abnormal plasma cell in bone marrow. Two different cytoplasmic mechanisms have been described [1] during NF-κB activation. First, the canonical pathway involving RelA:p50 dimer is essential for fast immune response. Second, a non-canonical pathway is activated in organogenesis, development and cell survival. It involves p52:RelB dimer. NF-κB nuclear activity and its effect in MM was studied using ChIP-seq experiments performed for each NF-κB subunit in MM cell line (MM.1S) in which the two NF-κB pathway are constitutively activated due to mutations inactivating *TRAF3*, a cytoplasmic inhibitor of the two NF-κB pathways of interest. Here, we present a complementary approach to characterize NF-κB nuclear activity in MM by combining different NGS datasets.

First, a ChIP-seq analysis is performed. This approach combines a quality assessment with fastqc, an alignment on hg19 human genome with bowtie2 and finally a peak calling using stringent threshold (p-value $<$ 1e-7 and FDR $< 1\%$) with macs2. We finally obtained 19,199 peaks for NF-κB subunit. In order to realize functional annotation of NF-κB peaks and MM.1S cell line, we used ChromHMM [2] with five histone marks to define 10 different chromatin states. We defined as *Promoters* all states showing enrichment for H3K4 trimethylation and H3K27 acetylation. *Enhancers* were defined as enrichment for H3K4 mono-methylation and H3K27 acetylation. To characterize how both pathway interacts in MM.1S, all κB regions are clustered using the k-mean method of seqMiner. It allows to identify different pattern of co-binding and correlate them to ATAC-seq enrichment. We also performed motif discovery using MEME-suite and motif enrichment determination with HOMER and ROC curves. Finally, we compared our results with DNAse-seq data of GM12878 (B cell lymphoma cell line) and ChromHMM data to compute differential enrichment of nucleosome free regions in this cell line.

Our analysis revealed that majority of NF-κB ChIP-seq peaks are located in promoters and in free accessible regions. Surprisingly, no κB motif enrichment was found for RelA and p52 subunit, but only in promoters binding RelB and in enhancers binding RelB and/or p50 subunit indicating the major role of this dimer for NF-κB specific binding. We also demonstrated that an interaction between the two NF-κB pathways exists in MM as it was shown in other malignancies [3]. Some differences on the changing of regulatory regions between those two cell lines shows the lack of enhancer location and κB motif enrichment in MM.1S cell line and implies specific mechanisms for this malignancies.

Acknowledgements

This work is supported by the "Fondation Française pour la Recherche contre le Myélome et les Gammapathies monoclonales". Authors thank Magali Devic, Elise Douillard, Emilie Maurenton, and Nathalie Roi for excellent technical expertise.

References

[1] Louis M Staudt. Oncogenic activation of NF-κB. Cold Spring Harbor perspectives in biology, 2(6):a000109, 2010.

[2] Jason Ernst, Pouya Kheradpour, Tarjei S Mikkelsen, Noam Shoresh, Lucas D Ward, Charles B Epstein, Xiaolan Zhang, Li Wang, Robbyn Issner, Michael Coyne, et al. Mapping and analysis of chromatin state dynamics in nine human cell types. Nature, 473(7345):43–49, 2011.

[3] Bo Zhao, Luis A Barrera, Ina Ersing, Bradford Willox, Stefanie CS Schmidt, Hannah Greenfeld, Hufeng Zhou, Sarah B Mollo, Tommy T Shi, Kaoru Takasaki, et al. The nf-κb genomic landscape in lymphoblastoid b cells. Cell reports, 8(5):1595–1606, 2014.

Efficient MinION data management

Laurent JOURDREN[1], Aurélien BIRER[1] , Lionel FERRATO-BERBERIAN[1], Sophie LEMOINE[1] and Stéphane LE CROM[1]

[1] École normale supérieure, PSL Research University, CNRS, Inserm, Institut de Biologie de l'École normale supérieure (IBENS), Plateforme Génomique, 75005 Paris, France

Corresponding Author: jourdren@biologie.ens.fr

Keywords: High throughput sequencing, Oxford Nanopore Technologies, MinION, Data management, Demultiplexing, Quality control

Summary

In the last 12 months, the throughput of the MinION [1], a third generation sequencer provided by Oxford Nanopore Technologies (ONT), has dramatically raised. Today, with the last R9.4 chemistry more than one million of long sequence reads can be sequenced and much more are expected by the end of the year with the frequent updates (every 2-3 months) of the flowcell chemistry.

ONT choose to create one raw file (FAST5 file) for each read produced by the MinION. Moreover the basecalling [2] of raw data is an intensive computational step as a neuronal network algorithm is used. Hence, MinION data management and quality control is a disk space and time consuming task. Dealing manually with the various input and output file formats from sequencer software requires many hours/days to generate FASTQ files and QC reports.

In this poster, we present the best practices to deal with MinION data from raw data to QC reports in an efficient manner. We now plan to create a new tool in order to automatically handle data transfer, read demultiplexing conversion and quality control once a sequencing run has been finished as we previously have done with Aozan [3] for Illumina sequencing.

Acknowledgements

This work was supported by the France Génomique national infrastructure, funded as part of the "Investissements d'Avenir" program managed by the Agence Nationale de la Recherche (contract ANR-10-INBS-09).

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/nanoporetech/nanonet
[3] Sandrine Perrin *et al.* Aozan: an automated post-sequencing data-processing pipeline. *Bioinformatics*, 2017.

Mise en place d'un pipeline de contrôle de la qualité de runs MinIon.

Lionel Ferrato-Berberian[1], Aurélien Birer[1], Ammara Mohammad[1], Corinne Blugeon[1], Fanny Coulpier[1], Stéphane Le Crom[1], Laurent Jourdren[1], Sophie Lemoine[1]

[1] École normale supérieure, PSL Research University, CNRS, Inserm, Institut de Biologie de l'École normale supérieure (IBENS), Plateforme Génomique, 75005 Paris, France.

Corresponding author : ferrato@biologie.ens.fr

Keywords : MinIon, nanopore, pipeline, analyse, QC

La technologie oxford nanopore MinIon est une technologie de séquençage qui permet l'obtention de lectures plus longues que la technologie illumina.

La technologie de séquençage ne repose pas sur des techniques d'imagerie comme illumina mais consiste à mesurer la différence de potentiel détecté lors du passage du brin d'ADN dans les pores.

On obtient alors des signaux électriques qui vont être convertis en lecture. Les données sont produites sous un format fast5. Les outils de suivis de runs existants ne sont plus adaptés ni aux métriques à suivre ni au format de fichier généré d'où la nécessité de développer des outils de contrôle qualité spécifiques.

Une bibliographie des outils d'analyses de la qualité des runs existants (poRe[1], poretools[2], ioniser[3], minotour[4]) a été réalisée pour prendre les meilleurs aspects de chaque outil afin de les combiner dans un nouveau pipeline d'analyse du run.

Ce poster présente le pipeline et les résultats produits par celui-ci de façon à permettre l'évaluation optimale des résultats d'un run MinIon. Notre pipeline a été développé et appliqué sur des données de runs de RNASeq MinIon comparant les transcrits WT et KO Egr2[5].

Bibliographie

1 : https://github.com/mw55309/poRe_docs
2 : https://poretools.readthedocs.io/en/latest/
3 : http://bioconductor.org/packages/release/bioc/html/IONiseR.html
4 : http://minotour.nottingham.ac.uk/
5 : Topilko, P. et al. Krox-20 controls myelination in the peripheral nervous system. Nature 371, 796–799 (1994).

Impact et évolution de la correction d'erreur sur des lectures longues issues de séquençage MinION Oxford Nanopore dans un contexte transcriptomique

Lionel FERRATO-BERBERIAN[1], Aurélien BIRER[1], Stéphane LE CROM[1], Laurent JOURDREN[1] and Sophie LEMOINE[1]

[1] École normale supérieure, PSL Research University, CNRS, Inserm, Institut de Biologie de l'École normale supérieure (IBENS), Plateforme Génomique, 75005 Paris, France.

Corresponding author: slemoine@biologie.ens.fr

Les technologies de séquençage de 3ème génération comme celle proposée par Oxford Nanopore (ONT), permettent d'obtenir des lectures longues de plusieurs milliers de kilobases. Ces lectures longues ouvrent des champs nouveaux dans un contexte transcriptomique comme l'accès direct aux transcrits alternatifs et aux modifications post-transcriptionnelles.

Contrairement aux séquences issues des technologies de 2ème génération, les séquences longues comportent beaucoup d'erreurs, principalement des insertions-délétions. Ces erreurs peuvent être corrigées en utilisant des méthodes hybrides, les séquences courtes de type Illumina viennent corriger les lectures longues, ou des méthodes non-hybrides, basées sur une couverture importante du transcriptome. Cette étape de correction alourdit l'analyse des données menant à l'obtention d'un transcriptome de référence.

La qualité des séquences en sortie de MinION s'améliorant à chaque version de chimie, de flowcell, de logiciel d'appel de base, nous nous proposons de tester l'impact de la correction d'erreur sur différente version de séquençage MinION de façon à envisager l'opportunité ou non de la correction des lectures en RNA-Seq.

Toullig: New pipeline for nanopore data analysis

Aurélien BIRER[1], Lionel FERRATO-BERBERIAN[1], Stéphane LE CROM[1], 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.

Corresponding Author: birer@biologie.ens.fr

keywords: High throughput sequencing, Oxford Nanopore Technologies, MinION, RNA-Seq, Mapping, Quality control

Summary

Since the release of Oxford Nanopore Technologies (*ONT*) MinION sequencers in 2014 [1] the number of reads produced with this new sequencing technology is still increasing. All the protocols remain in very active development (ONT provides updates of its chemistry and bioinformatics tools every 2-3 months). Hence the bioinformatical tools must be up to date throughout the development of MinION technology.

The IBENS (Institut de Biologie de l'École normale supérieure) genomic facility is currently developing a new data analysis workflow for RNA-Seq experiments using ONT sequencing output: *Toullig*. This pipeline is based on Eoulsan [2] and its bundled RNA-Seq pipeline for the Illumina reads.

The final goals of Toullig is to perform differential expression analysis from ONT long reads and produce a reference transcriptome by combining data from both Illumina and ONT technologies.

In this poster, we present our work on the Toullig pipeline with a focus on the long read mapping [3,4,5,6] and mapping quality control [7,8] steps.

The new Eoulsan modules for Toullig and the toolbox for manipulating ONT data are available on GitHub [9].

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] Jourdren, L., Bernard, M., Dillies, M.-A. & Le Crom, S. Eoulsan: a cloud computing-based framework facilitating high throughput sequencing analyses. *Bioinformatics* 28, 1542–1543 (2012).
[3] https://github.com/lh3/bwa.
[4] Sović, I. *et al.* Fast and sensitive mapping of nanopore sequencing reads with GraphMap. *Nat. Commun.* 7, 11307 (2016).
[5] Wu, T. D. & Watanabe, C. K. GMAP: a genomic mapping and alignment program for mRNA and EST sequences. *Bioinforma. Oxf. Engl.* 21, 1859–1875 (2005).
[6] Kiełbasa SM, Wan R, Sato K, Horton P, Frith MC. Adaptive seeds tame genomic sequence comparison. Genome Res. 2011 21(3):487-93.
[7] Weirather, J. L. *et al.* Comprehensive comparison of Pacific Biosciences and Oxford Nanopore Technologies and their applications to transcriptome analysis. *F1000Research* 6, 100 (2017).
[8] https://github.com/s-andrews/BamQC.
[9] https://github.com/GenomicParisCentre/toullig.

Extreme phenotypes define epigenetic and metabolic myeloid signatures in cardiovascular diseases.

D. Seyres [1,2,3] J.L. Lambourne [1,2], P. Kirk [4], A. Cabassi [4], F. Burden [1,2], R. Kreuzhuber [1,2,5], S. Farrow [1,2], C. Kempster [1,2], H. Mckinney [1,2], A. Park [6], D. Savage [7], J. Griffin [4], O. Stegle [5], S. Richardson [4], K. Downes [1,2], W.H. Ouwehand [1,2,8,9,10], M. Frontini [1,2,10]

[1] Department of Haematology, University of Cambridge, Cambridge Biomedical Campus, Cambridge, CB2 0PT, UK
2 National Health Service Blood and Transplant (NHSBT), Cambridge Biomedical Campus, Cambridge, CB2 0PT, UK
3 NIHR BioResource-Rare Diseases, University of Cambridge, Cambridge Biomedical Campus, Cambridge, CB2 0QQ, UK
4 Medical Research Council Biostatistics Unit, University of Cambridge, Forvie Site, Cambridge Biomedical Campus, Cambridge, CB2 0SR, UK
5 European Molecular Biology Laboratory, European Bioinformatics Institute, Wellcome Genome Campus, Hinxton, Cambridge, CB10 1SD, UK
6 Department of Clinical Biochemistry, Box 232 Addenbrooke's Hospital Hills Road Cambridge CB2 0QQ, UK
7 University of Cambridge Metabolic Research Laboratories, Wellcome Trust-MRC Institute of Metabolic Science, Box 289, Addenbrooke's Hospital, Cambridge CB2 0QQ, UK
8 The National Institute for Health Research (NIHR) Blood and Transplant Unit in Donor Health and Genomics at the University of Cambridge, University of Cambridge, Strangeways Research Laboratory, Cambridge, CB1 8RN, UK
9 Department of Human Genetics, The Wellcome Trust Sanger Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge, CB10 1SA, UK
10 BHF Centre of Excellence, Division of Cardiovascular Medicine, Addenbrooke's Hospital, Cambridge Biomedical Campus, Cambridge, CB2 0QQ, UK

Corresponding Authors: ds777@medschl.cam.ac.uk, mf471@medschl.cam.ac.uk

Almost 6 million people are affected by cardiovascular disease (CVD) and 26% of deaths are attributed to CVD in the UK each year. Thus, early detection of CVD onset is critical to improve quality of life and to decrease the economic burden on healthcare providers. The identification of predictive biomarkers and interpretable disease signatures by combining data obtained with different high-throughput omics is the ultimate step towards reconstruction and analysis of complex multi-dimensional diseases, enabling deeper mechanistic and medical insight. To this end, we collected data on 169 blood donors and 22 patients representing three different diseases with high CVD risk. For each individual, monocytes and neutrophils were isolated and underwent whole genome sequencing, ChIP-sequencing for histone modifications representing regulatory elements (H3K4me1, H3K27ac), RNA-sequencing and DNA methylation analysis. Additionally, plasma metabolites and lipids were quantified in all individuals.

We are currently integrating the different data types, working to select relevant covariates using penalized likelihood approaches, followed by clustering via variational Bayes mixture models and other clustering methods [1,2,3] according to continuous responses represented by clinical values associated with a high CVD risk factor. Genetic contribution to a high CVD risk factor is analyzed by performing quantitative trait loci mapping using histone modifications, methylation, transcripts expression, lipidomic and metabolic data.

Our aim is to identify metabolic and/or epigenetic signatures that could be applied for the detection of CVD from an early onset.

References

[1] Witten & Tibshirani. A framework for feature selection in clustering. *Journal of the American Statistical Association*, 105(490), 713–726). 2010
[2] Liverani et al. PReMiuM: An R Package for Profile Regression Mixture Models Using Dirichlet Processes. Journal of Statistical Software, 64(7)). 2015
[3] Shen et al. Integrative clustering of multiple genomic data types using a joint latent variable model with application to breast and lung cancer subtype analysis. *Bioinformatics* 25 (22): 2906-2912. 2009

Analyse et co-développement d'outils de bioinformatique destinés au traitement de données NGS issues de métagénomique virale

Emilie DELPUECH[1], Guillaume CROVILLE[2], Jean-Luc GUERIN [2], Christophe KLOPP[1] and Sarah MAMAN[1]

[1] Plateforme Bioinformatique Toulouse Midi-Pyrénées, UBIA, INRA, Chemin de Borde Rouge, 31326 Castanet-Tolosan, France

[2] IHAP, Université de Toulouse, INRA, ENVT, 23 Chemin des Capelles, 31300 Toulouse, France

Corresponding Author: emilie.delpuech@inra.fr

Les techniques traditionnelles de détection des agents infectieux (PCR, recherche d'antigènes ou d'anticorps via la technique ELISA) par mise en culture (sur cellules ou œufs embryonnés), sont à la fois très lourdes, coûteuses et souvent pas suffisamment sensibles. Des approches dites « sans *a priori* » sont désormais susceptibles de les supplanter.

Les nouvelles générations de séquençage (NGS) sont devenues essentielles pour l'étude de l'identité et de la variabilité des génomes microbiens. Les techniques de type MiSeq (Illumina) et MinIon (Oxford Nanopore) sont aujourd'hui complémentaires, de par la nature et le volume de séquences générées. Une première étape d'étude bibliographique a permis de découvrir plusieurs *pipeline*s répondant à la problématique de détection de virus. Trois d'entre eux ont été étudiés plus spécifiquement : Kraken [1], Truffle [2] et VIP [3].

Le *pipeline* Kraken [1] utilise une méthode d'alignement exact de K-mers pour réaliser une affiliation taxonomique des séquences d'agents pathogènes présentes dans des données NGS. Les K-mers sont recherchés dans la base de données Kraken. La correspondance exacte des 31-mers rend le *pipeline* stringent. Est-ce un facteur rédhibitoire pour détecter spécifiquement les séquences virales de données NGS de métagénomique virale ?

La détection de virus peut être réalisée à l'aide de l'outil Truffle [2] qui utilise des e-sondes représentantes des virus recherchés. Par conséquent, cet outil nécessite la connaissance des virus potentiellement présents dans les jeux de données. Ces e-sondes sont générées et par la suite, recherchées par alignement dans les données NGS. Un score est calculé pour chaque e-sonde. Le point fort de ce *pipeline* est l'obtention de scores représentatifs de la présence du virus, via une analyse statistique.

Le *pipeline* VIP [3] utilise quant à lui, une méthode d'assemblage *de novo*. Après un contrôle qualité et la suppression des séquences hôtes éventuelles, les lectures sont ensuite assemblées en contigs puis assignées via un alignement nucléotidique contre des bases de données virales spécifiques (ViPR/IRD), ou contre les bases de données virales du NCBI. A partir de cette assignation un assemblage *de novo* permet de reconstruire le génome des virus détectés. En complément, une analyse phylogénétique est proposée. Ce *pipeline* développé pour la détection de virus humains dans le cadre clinique, peut être appliqué à la détection de virus d'intérêt vétérinaire.

Cette étude a permis une prise en main de ces *pipeline*s pour rendre chacun d'eux compatible, avec des jeux de données provenant de séquençages de tissus animaux. Les *pipeline*s ont été testés et comparés afin d'optimiser leurs paramétrages. D'autre part nous réalisons une étude de faisabilité de l'intégration du *pipeline* VIP à la plateforme web d'analyse de données, Galaxy [4]. La mise à disposition d'un pipeline d'identification et de découverte de virus, est importante pour le diagnostic microbiologique, la surveillance de virus d'intérêt en santé publique et plus largement, la découverte de nouveaux virus.

References

[1] D.E. Wood, S.L. Salzberg, Kraken: ultrafast metagenomic sequence classification using exact alignments, Genome Biol. 15 (2014) R46. doi:10.1186/gb-2014-15-3-r46.

[2] M. Visser, J.T. Burger, H.J. Maree, Targeted virus detection in next-generation sequencing data using an automated e-probe based approach, Virology. 495 (2016) 122–128. doi:10.1016/j.virol.2016.05.008.

[3] Y. Li, H. Wang, K. Nie, C. Zhang, Y. Zhang, J. Wang, P. Niu, X. Ma, VIP: an integrated pipeline for metagenomics of virus identification and discovery, Sci. Rep. 6 (2016) 23774. doi:10.1038/srep23774.

[4] B. Giardine, C. Riemer, R.C. Hardison, R. Burhans, L. Elnitski, P. Shah, Y. Zhang, D. Blankenberg, I. Albert, J. Taylor, W. Miller, W.J. Kent, A. Nekrutenko, Galaxy: A platform for interactive large-scale genome analysis, Genome Res. 15 (2005) 1451–1455. doi:10.1101/gr.4086505.

3-SMART: Bioinformatic analysis of intronic polyadenylation regulation

Mandy CADIX[1,2], Iris TANAKA[2], Pierre GESTRAUD[1], Marine SÉJOURNÉ[1,2], Stéphan VAGNER[2], Nicolas SERVANT[1] and Martin DUTERTRE[2]

[1] Bioinformatics and Computational Systems Biology of Cancer (U900) – Institut Curie - PSL Research University - MINES ParisTech - Inserm, 26 rue d'Ulm, 75248 Paris cedex 05, France

[2] CNRS UMR3348 – Institut Curie – PSL Research University, Centre Universitaire Bâtiments 110, 91405 Orsay, France

Corresponding Author: mandy.cadix@curie.fr

The 3' end maturation of precursor mRNAs includes 3' end cleavage and addition of the polyadenylated tail. In addition, alternative polyadenylation (APA) occurs in about half of the genes in mammals. There are two types of alternative polyadenylation. The first one is located upstream of the last exon of the gene and is called intronic polyadenylation (IPA). The second is located within the last exon of the gene and is called polyadenylation in tandem. APA is widely regulated in oncogenic transformation and cell response to genotoxic (DNA-damaging) agents. 3'-seq (high-throughput sequencing of 3'-ends of polyadenylated transcripts) is a particular type of RNA sequencing technique. This methods allows sequencing the 3' end of the transcript upstream of the polyadenylated tail, in order to evaluate APA regulation.

The aim of this work was to develop a bioinformatics pipeline for differential 3'-seq analysis, called 3-SMART (3'-seq Mapping Annotation and Regulation Tool). Our pipeline includes 6 steps : 1/ Quality control and trimming of reads ; 2/ Reads mapping to genome ; 3/ Identification of peaks ; 4/ Annotation of peaks (IPA or last exon of genes) ; 5/ Filtering out of artefactual peaks corresponding to internal priming of oligo(dT) primers during reverse-transcription ; and 6/ Differential analysis of each IPA between two conditions. We applied this approach to a set of experimental data obtained in our lab to analyze IPA regulation in lung cancer cell response to cisplatin, a genotoxic anticancer drug. We found that cisplatin up-regulates IPA in many genes. The 3-SMART pipeline was implemented in bash and required a linux system. The pipeline is available at GitHub (https://github.com/InstitutCurie/3-SMART).

Developing and sharing reproducible bioinformatics pipelines: best practices

Yohann LELIEVRE[1], Audrey BIHOUÉE[2], Eric CHARPENTIER[2], Alban GAIGNARD[2,4], Simon SOUCHET[3] and Damien VINTACHE[1]

[1] LS2N, UMR CNRS 6004, IMT Atlantique, ECN, Université de Nantes, Nantes, France.
[2] l'institut du thorax, INSERM, CNRS, Université de Nantes, Nantes, France.
[3] Angers Academic Hospital, CHU d'Angers, France
[4] Nantes Academic Hospital, CHU de Nantes, France

Corresponding author: Yohann.Lelievre@univ-nantes.fr

1 Introduction

Life-sciences are nowadays conducted in multi-disciplinary and multi-centric studies. In this context, the same software components must be deployed in multiple environments for reproducibility and scalability issues. In addition, data analysis pipelines are usually composed of multiple components, continuously evolving, which leads to maintenance and long-term support challenges. To promote FAIR [1] principles, providing controlled software environments becomes mandatory. We propose a set of best practices taking advantage of proven or promising tools: Git, Conda, SnakeMake[1], Jenkins and Docker.

2 Motivations

Bio-informaticians and software developers need to build data analysis pipelines in controlled environments to ensure long-term re-execution and better reproducibility. From an end-user point of view, typically a biologist, data analysis pipelines should be automatically installable in a local or dedicated computing infrastructure, including any software or data dependency. Pipelines should be launched in three steps: i) environment setup/activation, ii) parameters tuning, and iii) pipeline execution.

3 Approach and Results

The BiRD pipeline registry results from applying these guidelines in the context of Exome sequencing and RNAseq (variant calling, differential gene expression, gene fusion detection, single-cell). These pipelines are described in a GitLab web portal. GitLab allows i) to document the pipeline an its usage, and ii) to host and version the associated source code. To ease installation and dependency management, we packaged and deployed the executable software components through the Conda package manager in a dedicated repository [2]. To assess their long-term re-execution, workflows and associated software environments are nightly assembled into minimal Docker images through a Jenkins continuous integration system.

4 Conclusion and perspectives

The best practices hereby proposed aim at promoting findable and accessible data analysis pipelines through web-based resources. This process allows to re-package and re-execute pipelines in the long run, and to adapt to continuously evolving environments. Our future works include two main directions: i) handling data resources as part of the pipeline distribution process (*e.g.* BioMaj), and ii) studying how to promote interoperability between multiple systems and infrastructures. To enhance trust for end-users and to encourage reuse, provenance metadata and controlled vocabularies (*e.g.* EDAM) offer interesting perspectives to associate produced/analyzed with large-scale bio-resource registries such as BioTools.

Acknowledgements

This work was supported by the BiRD core facility, the SyMeTRIC project and the GRIOTE project.

References

[1] Johannes Köster and Sven Rahmann. Snakemake - a scalable bioinformatics workflow engine. *Bioinformatics*, (28 (19)):2520–2522, 2012.

1. Findable - Accessible - Interoperable - Reusable
2. https://anaconda.org/BiRD

MICROSCOPE: an integrated platform for the Exploration and Curation of Microbrial Genomes

David VALLENET[1], Alexandra CALTEAU[1], Stéphane CRUVEILLER[1], Mathieu GACHET[1], Guillaume GAUTREAU[1], Adrien JOSSO[1], Aurélie LAJUS[1], Jordan LANGLOIS[1], Jonathan MERCIER[1], Hugo PEREIRA[1], Rémi PLANEL[1], Johan ROLLIN[1], David ROCHE[1], Zoé ROUY[1], Claudine MÉDIGUE[1]
CEA/Genoscope/LABGeM, Université d'Evry Val-d'Essonne, CNRS-UMR 8030, Université Paris-Saclay, 2 rue Gaston Crémieux, 91057, Evry, France

Corresponding author: vallenet@genoscope.cns.fr

1 Introduction

The analysis of genomes from NGS platforms needs to be automated and fully integrated. However, maintaining consistency and accuracy in annotation is a challenging task because millions of proteins in databanks are not assigned reliable functions.

The LABGeM team of the Genoscope sequencing center focuses its research activities on the development and application of new methods for genome analysis. These tools are then made available through MicroScope (http://www.genoscope.cns.fr/agc/microscope) [1], an integrated platform dedicated to microbial genome annotation and comparative analysis, which is being developed in our group since 2004.

2 Methods

The resource provides data from complete and ongoing genome projects together with post-genomic experiments (i.e. transcriptomics, re-sequencing of evolved strains) allowing users to improve the understanding of gene functions. We will present an overview of the MicroScope analysis pipelines and illustrate the use of several new functionalities in the context of data discovery and expert annotation, which concern:

- comparative genomics with synteny computations and pan-genome analyses,

- the prediction of virulence and antimicrobial resistance genes,

- the detection and annotation of genomic regions of interest, like, secretion systems, integrons and secondary metabolite biosynthesis gene clusters,

- and metabolic network reconstruction assisted by the GROOLS expert system (https://github.com/grools) [2].

3 Conclusions

To date, MicroScope contains data for about 7,000 microbial genomes, part of which are manually curated and maintained by microbiologists ($>$ 3,200 personal accounts in March 2017). The platform enables collaborative work in a rich comparative genomic context and improves community-based curation efforts.

References

[1] David Vallenet *et al.* Microscope in 2017: an expanding and evolving integrated resource for community expertise of microbial genomes. *Nucleic Acids Res.*, (45(D1)):D517–D528, 2017.

[2] Jonathan Mercier *et al.* Grools: reactive graph reasoning for genome annotation through biological processes. *bioRxiv*, 2017.

NETSYN : NETwork SYNteny, a new tool to help functional annotation

Benjamin VIART[1], Karine BASTARD[1], Guillaume REBOUL[2] z, Hugo PEREIRA[1], Remi PLANEL[1], Mark STAM[1], Claudine MEDIGUE[1] and David VALLENET[1]

[1] LABGeM, CEA, Genoscope, Université d'Evry, CNRS-UMR8030, Université Paris-Saclay, 2 rue Gaston Cremieux, 91057 Evry, France

[2] Unité d'écologie systématique et évolution, CNRS UMR 8079, Université Paris Sud

Corresponding author: bviart@genoscope.cns.fr

1 Abstract

About 25% of the protein families collected in the Pfam database are annotated with unknown functions [1]. Besides, in databanks in which proteins are mainly annotated based on sequence similarity only (e.g. TrEMBL, NR), the level of misannotation reaches about 60% for superfamilies [2]. To detect possible errors or to suggest hypothetical functions for these proteins, we developed a bioinformatics approach, called NetSyn, based on genomic neighborhood.

NetSyn aims at exploring conserved genomic contexts (i.e. syntenies) to classify proteins within a family. It consists of (*i*) a pairwise comparison of gene chromosomal organization to find conserved syntenies, (*ii*) the representation of the results using a non-oriented weighted graph where nodes represent proteins and edges the average score of the number of genes involved in the synteny, and (*iii*) nodes are are clustered using the weighted Markov Cluster Algorithm [3].Each cluster is supposed to gather iso-functional proteins. The software can be used either to study multi-functional enzyme families or to find interactions between several families.

As a proof of concept, NetSyn was tested on two families. In the β-keto acid cleavage enzymes family, 7 profiles of active sites have been described, each one has been linked experimentally with specific enzymatic activities [4]. NetSyn was able to retrieve the proteins associated with the 7 known activities and to disclose 4 other clusters of proteins that might have new functions. In the amidinotransferases family [5], proteins annotated as arginine deiminase were split in two different clusters. One of them probably contains proteins that might not be correctly annotated as these proteins do not share similar genomic context with characterized arginine deiminase. Our method points out another cluster in which half the proteins are annotated as "Unknown function". Because they are highly connected with enzymes experimentally characterized as dimethylarginine dimethylaminohydrolase, these proteins might share the same enzymatic activities. The program is available to the scientific community on GitHub.

References

[1] R. Mudgal, S. Sandhya, N. Chandra, and N. Srinivasan. De-DUFing the DUFs: Deciphering distant evolutionary relationships of Domains of Unknown Function using sensitive homology detection methods. *Biol. Direct*, 10:38, Jul 2015.

[2] A. M. Schnoes, S. D. Brown, I. Dodevski, and P. C. Babbitt. Annotation error in public databases: misannotation of molecular function in enzyme superfamilies. *PLoS Comput. Biol.*, 5(12):e1000605, Dec 2009.

[3] S. van Dongen. A cluster algorithm for graphs. *Technical Report INS R0010, National Research Institute for Mathematics and Computer Science in the Netherlands*, May 2000.

[4] K. Bastard, A. A. Smith, C. Vergne-Vaxelaire, A. Perret, A. Zaparucha, R. De Melo-Minardi, A. Mariage, M. Boutard, A. Debard, C. Lechaplais, C. Pelle, V. Pellouin, N. Perchat, J. L. Petit, A. Kreimeyer, C. Medigue, J. Weissenbach, F. Artiguenave, V. De Berardinis, D. Vallenet, and M. Salanoubat. Revealing the hidden functional diversity of an enzyme family. *Nat. Chem. Biol.*, 10(1):42–49, Jan 2014.

[5] H. Shirai, T. L. Blundell, and K. Mizuguchi. A novel superfamily of enzymes that catalyze the modification of guanidino groups. *Trends Biochem. Sci.*, 26(8):465–468, Aug 2001.

Stratégie d'assemblage de génomes en cellule unique de protistes marins dans le cadre du projet Tara Oceans

Léo D' AGATA[1], Yoann SEELEUTHNER[1], Julie POULAIN[1], Patrick WINCKER[2], and Jean-Marc AURY[1]

[1] Commissariat à l'Energie Atomique (CEA), Institut de Génomique (IG), Genoscope, F-92057 Evry, France.

[2] Genoscope, CEA, CNRS UMR 8030, Université d'Evry, France.

Corresponding Author: ldagata@genoscope.cns.fr

Le plancton des océans est responsable de plus de la moitié de l'oxygène produit sur la planète. Il n'en demeure pas moins très peu étudié et l'on connait encore mal l'impact que pourrait avoir le réchauffement climatique et la pollution sur ce dernier. Son rôle dans l'absorption du CO2 et la régulation du climat pourrait en effet être directement affecté. C'est dans le but de répondre à ces questions que l'expédition Tara, à laquelle le Genoscope participe, fut lancée entre 2009 et 2013 sous le nom de *Tara* Oceans [1].

Un des objectifs est l'étude précise des génomes de protistes marins. Ces derniers n'étant pour la plupart pas cultivables en laboratoire, la solution mise en place fut de recourir au séquençage en cellule unique. Le protocole de ce dernier consistait à extraire l'ADN d'une seule cellule, à l'amplifier par MDA [2] et enfin à le séquencer. Les étapes d'extraction et d'amplification génèrent toutefois de nombreux problèmes rendant l'assemblage de leur génome difficile. Certaines régions peuvent en effet être non capturées lors de l'extraction de l'ADN tandis que d'autres ne sont pas amplifiées. L'amplification appliquée à ces protistes génère de plus une couverture verticale très irrégulière.

Des outils adaptés sont donc nécessaires pour assembler ces génomes. Un workflow spécifique a ainsi été mis en place au Genoscope pour traiter ces données. Une idée proposée par ce workflow consiste notamment à se servir de la synergie de plusieurs cellules pour résoudre au mieux les problèmes de régions non couvertes. Cette technique dite de « co-assemblage » nécessite néanmoins de savoir à l'avance si ces cellules appartiennent au même organisme. L'ADN ribosomique 18S (gène marqueur chez les eucaryotes) peut apporter un début de réponse. Cette séquence n'est cependant pas toujours très résolutive et ne reflète pas totalement la diversité des génomes.

Pour répondre à cette problématique nous avons utilisé différentes méthodes et outils associés. Deux méthodes principales ont été retenues. La première consiste à se servir de la technique des graphs de De Bruijn colorés sur un premier co-assemblage [3]. Chaque couleur correspondant à une cellule, il nous est possible d'observer par quelles cellules sont recouverts les contigs. Les régions génomiques partagées entre les cellules nous informent ainsi sur leur proximité ou leur éloignement. La deuxième méthode consiste quant à elle à découper les lectures en mots chevauchants de longueur k (k-mers) et à mesurer la proportion de ces mots qui sont partagés entre deux cellules séquencées de façon indépendante [4]. Plus le taux de k-mers partagés par deux cellules est élevé, plus on peut supposer que celles-ci appartiennent à la même espèce. L'utilisation conjointe de ces méthodes nous aide donc à choisir les cellules pouvant être ou non co-assemblées.

References

[1] Eric Karsenti, Silvia G Acinas, et al. A holistic approach to marine eco-systems biology. *PLoS Biol* 9(10): e1001177, 2011.

[2] Frank B Dean, Seiyu Hosono, et al. Comprehensive human genome amplification using multiple displacement amplification. *Proc Natl Acad Sci USA* 99(8):5261-5266, 2002.

[3] Narjes S Movahedi, Mallory Embree, et al. Efficient Synergistic Single-Cell Genome Assembly. *Front Bioeng Biotechnol* 4: 42, 2016.

[4] Gaëtan Benoit, Pierre Peterlongo, et al. Multiple comparative metagenomics using multiset k-mer counting. *PeerJ Computer Science* 2:e94, 2016.

Assessing the functional impact of genomic alterations using proteogenomics

Georges Bedran[1,2,3], Yves Vandenbrouck[1], Eric Bonnet[2], Jean-François Deleuze[2],

Delphine Pflieger[1,4*] and Christophe Battail[1,2*]

[1] CEA/DRF/BIG/BGE/EDYP, 17 rue des Martyrs, 38054, Grenoble, France
[2] CEA/DRF/JACOB/CNRGH, 2 rue Gaston Crémieux, 91057, Evry, France
[3] UFR Sciences et Techniques - Université de Rouen, Place Emile Blondel, 76821, Mont-Saint-Aignan, France
[4] CNRS, 25 rue des Martyrs, Grenoble, 38042, France
[*] Co-senior authors

Corresponding Author: georges.bedran@etu.univ-rouen.fr

Proteomic data is obtained using a combination of liquid chromatography and tandem mass spectrometry (MS/MS) where peptides are most commonly identified by matching MS/MS spectra against theoretical spectra of all candidate peptides represented in a generalist protein sequence reference database[1] . The limitation of this approach is that variant peptides are missing from this reference database, thus they cannot be detected.

Proteogenomics[2] is an alternative approach where the reference database is replaced by a customized protein sequence database generated using genomic information extracted from RNA-seq data[3]. This strategy allows the detection of novel SAAVs, INDELs and splice junctions.

Using exploratory proteogenomics on a colorectal cancer cell line, we identified over 134 Single Amino Acid Variants (SAAV), from which 88 were also found by CPTAC colorectal cancer study[4]. However, in such analysis a significant number of alterations found by RNA-seq are not detected at the protein level (3605 non-synonymous mutations); therefore, we are aiming to address them using a targeted approach [5].

Our perspective is to develop a bioinformatic methodology capable of assessing the impact of the genetic alterations on the proteome in a cancer context.

Keywords

Proteogenomics, RNA-sequencing, colorectal cancer, SAAVs (single amino acid variants), splice junctions, INDELs (insertions / deletions), exploratory proteomics, targeted proteomics.

References

1. Nesvizhskii AI. A survey of computational methods and error rate estimation procedures for peptide and protein identification in shotgun proteomics. J. Proteomics. 2010;73:2092–123.

2. Nesvizhskii AI. Proteogenomics: concepts, applications and computational strategies. Nat Meth. 2014;11:1114–25.

3. Wang X, Zhang B. customProDB: an R package to generate customized protein databases from RNA-Seq data for proteomics search. Bioinformatics [Internet]. 2013; Available from: http://bioinformatics.oxfordjournals.org/content/early/2013/09/20/bioinformatics.btt543.abstract

4. Zhang B, Wang J, Wang X, Zhu J, Liu Q, Shi Z, et al. Proteogenomic characterization of human colon and rectal cancer. Nature. 2014;513:382–7.

5. Yeom J, Kabir MH, Lim B, Ahn H-S, Kim S-Y, Lee C. A proteogenomic approach for protein-level evidence of genomic variants in cancer cells. Sci. Rep. 2016;6:35305.

Deciphering early cell fate decision by Single Cell RNA-Seq and DGE-Seq

Dimitri MEISTERMANN[1,2], Yohann LELIEVRE[2], Eric CHARPENTIER[3], Stéphanie KILENS[1], Thomas FREOUR[1,4], Jérémie BOURDON[2] and Laurent DAVID[1,5]

[1] CRTI, INSERM UMR1064, Université de Nantes, F-44000, Nantes, France
[2] LS2N, UMR 6004, Université de Nantes, F-44000, Nantes, France
[3] l'institut du thorax, INSERM UMR1087, CNRS UMR6291, Université de Nantes, F-44000, Nantes, France
[4] CHU de Nantes, Service de Biologie de la Reproduction, F-44000, Nantes, France
[5] iPSC core facility, INSERM UMS 016, SFR Francois Bonamy, F-44000, Nantes, France;
Corresponding author: dimitri.meistermann@univ-nantes.fr

The major goal of our laboratory is to understand and quantify human preimplantation development, from fertilization to implantation in the uterus, and to predict its outcome. Specifically, we aim at deciphering the molecular mechanism driving cell fate during this first step of our existence. Understanding human preimplantation is therefore critical to improve assisted reproductive technologies (ART) and broaden the use of human pluripotent stem cells in regenerative medicine. It is during this timeframe that embryonic cells make their first choice of cellular fate, moving from one totipotent cell in the zygote to an embryo stratified by three cell types in the mature blastocyst. Moreover, the main objective of in vitro fertilization is to support the development of the zygote into a blastocyst, before transfer in infertile patients. To discover how early cell fate specification is regulated, our team develops several strategies to model embryos. To reach this goal, we employed two transcriptomics approaches: single-cell RNA-Seq and DGE RNA-Seq.

Firstly, we used *Single-Cell RNA-Seq* to identify unknown subpopulation of preimplantation embryo cells and associated genes network. For this purpose, we designed a single-cell analysis pipeline written with *snakemake*, a new and handy workflow tool. Biological data in single-cell are highly heterogenous and require unsupervised approaches. Thus, the pipeline answer to these constraints and includes complete analysis from FASTQ to identification of cell subpopulations with dimension reduction methods. We also analyzed cell trajectories with pseudo-time methods (*Slingshot*, *Monocle2* [2]) to pinpoint critical moment of cell fate decisions. Then, we determined specific gene networks associated with subpopulations, embryo lineage or embryo stage with Weighted Gene Correlation Analysis (*WGCNA* [3]). We observed that WGCNA is a particularly suitable tool for single-cell analysis, giving coherent results from noisy and heterogenous data. Here we will present the pipeline and show how it has solved the problem of heterogeneity and lack of biological material inherent to human embryo. Secondly, a novel, cost-efficient RNA-Seq method, digital expression RNA-Seq (*DGE-Seq*) was employed to sequence human induced naive pluripotent stem cells (*hiNPSC*), a recently characterized state of pluripotency in human [4]. As a main result of this work, we show that hiNPSC have a more similar metabolism to preimplantation embryo than usual hIPS cells. This makes hiNPSC an appropriate model of preimplantation development.

Here we present methods of computational biology that guided our research strategy, and promoted the use of transcriptomics in Nantes. Both approaches contributed significantly to our understanding of preimplantation development, opening new avenues of research in the fields of ART and regenerative medicine fields.

Acknowledgements

This work was supported by the BiRD bioinformatics facility.

References

[1] J. Rossant et P. P. L. Tam, « New Insights into Early Human Development: Lessons for Stem Cell Derivation and Differentiation », *Cell Stem Cell*, vol. 20, nº 1, p. 18-28, janv. 2017.
[2] C. Trapnell *et al.*, « The dynamics and regulators of cell fate decisions are revealed by pseudotemporal ordering of single cells », *Nat. Biotechnol.*, vol. 32, nº 4, p. 381-386, mars 2014.
[3] P. Langfelder et S. Horvath, « WGCNA: an R package for weighted correlation network analysis », *BMC Bioinformatics*, vol. 9, p. 559, 2008.
[4] G. Guo *et al.*, « Naive Pluripotent Stem Cells Derived Directly from Isolated Cells of the Human Inner Cell Mass », *Stem Cell Rep.*, vol. 6, nº 4, p. 437-446, avr. 2016.

In search of W sex chromosome-specific sequences in the genome of the isopod crustacean *Armadillidium vulgare*

Mohamed Amine CHEBBI, Thomas BECKING, Isabelle GIRAUD, Bouziane MOUMEN,

Clément GILBERT, Jean PECCOUD and Richard CORDAUX

Laboratoire Ecologie et Biologie des interactions UMR CNRS 7267, Equipe Ecologie Evolution Symbiose, Université de Poitiers, Bât. B8, 5 rue Albert Turpin, TSA 51106, 86073 Poitiers Cedex 9 France

Corresponding Author: mohamed.amine.chebbi@univ-poitiers.fr

Abstract

In the isopod crustacean *Armadillidium vulgare*, genetic sex determination follows female heterogamety (ZZ males and ZW females). Z and W sex chromosomes show no apparent heteromorphy. WW female individuals are viable and genetic sex determination in the closely related species *Armadillidium nasatum* follows male heterogamety (XY males and XX females). These observations suggest that the evolution of *A. vulgare* sex chromosomes is at an incipient stage of the specialization of a pair of ancestral autosomes carrying sex determinants. To test this hypothesis and identify W-specific sequences, we sequenced and assembled the genome of *A. vulgare* by combining Illumina and PacBio sequencing technologies. The 1.7 Gb genome of *A. vulgare* is highly repeated, as simple repeats and transposable elements represent ~70% of the genome. To identify sex linked sequences, we obtained male and female Illumina sequencing reads independently and used two different bioinformatics approaches: Chromosome Quotient (CQ) and Y Genome screening (YGS) methods. These analyses confirmed the homomorphy of *A. vulgare* sex chromosomes and allowed us to identify candidate W-specific scaffolds.

Keywords:

female heterogamety, homomorphy, W-specific sequences, hybrid de novo genome assembly, highly repeated, Chromosome Quotient (CQ), Y Genome screening (YGS)

Exome sequencing to identify the molecular mechanism underlying chordomas pathogenesis

Zakia Tariq[1], Keltouma Driouch[1], Virginie Bernard[2], Ivan Bieche[1], Virginie Raynal[2], Sylvain Baulande[2], Hamid Mammar[3], Julien Masliah Planchon[14]

[1] Service de génétique, Institut Curie, 26 rue d'Ulm 75248, Paris, France.
[2] Next generation sequencing platform, ICGex, Institut Curie, Paris, France.
[3] Service radiothérapie, Institut Curie, Paris, France.
[4] INSERMU830, Département biologie des tumeurs, Institut Curie, Paris, France.

Corresponding Author: zakia.tariq@curie.fr

Chordomas are rare bone tumors, often observed in cranial, spinal and sacral sites. With a slow evolution, these tumors are clinically diagnosed at a late stage, decreasing patients' life expectancy [1]. However, these tumors remain poorly described. This project aims to identify and characterize somatic alterations involved in this pathology. To this end, a series of tumors were analyzed by means of high-throughput exome sequencing.

On Illumina platform, paired-end 100x100 exome-Seq was performed for eight primary tumors and matching germline DNAs obtained from blood samples. After quality control and mapping files cleaning, somatic variants search was performed. Therefore, three variant callers were used in parallel to increase the sensitivity and the confidence of predicted variations (SNV). After Annovar annotation, variants were considered as somatic when absent in the germline samples and if their reported frequency in 1000Genome database was below 0.1%. Finally, variants were validated by IGV visualization discarding sequencing and alignment errors. In another hand, copy number variations (CNV) were predicted using Facets and Sequenza tools, to determine putative copy number alterations.

The number of identified somatic mutations ranged from 12 to 32 per sample, except one tumor exhibiting 166 somatic variants. Among the mutated genes, known drivers, such as *KIT* and *PIK3CA*, were detected. Mutated genes highlighted recurrent pathways involved in biologic processes such as chromatin organization, and epigenetic modifications. With CNV analysis, we identified a chromothripsis phenomenon as formerly described [2]. In our study, such events were observed in 2 out of 8 tumors. Furthermore, we characterized a high frequency of *CDKN2A/B* homozygous deletion in 80% of samples; a much higher rate than previously reported [3]. Strikingly, the most extensively mutated sample, presented a homozygous deletion encompassing *MLH1* gene, suggesting a DNA mismatch repair deficiency in this tumor.

The integrated analysis of SNV and CNV profiles emphasized several molecular pathways involved in chordomas. In particular, the *MEFC2* gene, regulating bone differentiation, showed a biallelic inactivation i.e. a stop gain variation coupled with a loss of heterozygosity suggesting a complete loss of the protein function.

In conclusion, we observed different variations already described in the literature, but we also characterized novel putative genes involved in chordoma, not linked to the pathology yet. To validate our hypotheses, exome-Seq will be performed on a new sample series and RNA-Seq data will be generated to complete the study.

References

[1] Walcott BP, et al. Chordoma: Current concepts, management, and future directions. *The Lancet Oncology*, 13(2):e69-76, 2012.
[2] Stephens PJ, et al. Massive Genomic Rearrangement Acquired in a Single Catastrophic Event during Cancer Development. *Cell*, 144(1):27-40, 2011.
[3] Hallor KH, et al. Frequent deletion of the CDKN2A locus in chordoma: analysis of chromosomal imbalances using array comparative genomic hybridisation. *Br J Cancer*, 98: 434-42, 2008.

Revisiting cell lineage specification during male sex determination with single-cell RNA sequencing

Isabelle STÉVANT[1,2,3], Yasmine NEIRJINCK[1], Christelle BOREL[1], Jessica ESCOFFIER[1], Lee B. SMITH[4,5], Stylianos E. ANTONARAKIS[1,2], Emmanouil T. DERMITZAKIS[1,2,3], Serge NEF[1,2]

[1] Department of Genetic Medicine and Development, University of Geneva, 1211 Geneva, Switzerland;
[2] iGE3, Institute of Genetics and Genomics of Geneva, University of Geneva, 1211 Geneva, Switzerland;
[3] SIB, Swiss Institute of Bioinformatics, University of Geneva, 1211 Geneva, Switzerland;
[4] MRC Centre for Reproductive Health, University of Edinburgh, Edinburgh EH16 4TJ, UK;
[5] School of Environmental and Life Sciences, University of Newcastle, Callaghan, NSW 2308, Australia.

Corresponding Author: isabelle.stevant@unige.ch

Among the genetic characteristics that influence our physical identity all along our life, the most important is surely our sex. The ultimate goal of male and female sex differentiation and development is to provide organisms the necessary attributes for sexual reproduction. In mammals, the genetic sex of individuals is cast at fertilisation with the bring of an X or a Y chromosome from a spermatozoa to a bearing X oocyte. Phenotypic sex is only revealed during foetal development, when the gonads start to differentiate as ovaries or testes (around 6 embryonic weeks in human, and at embryonic day 11.5 in mouse). Consequently to the ovary or testis differentiation, the whole embryo will adopt the secondary sexual characteristics such as male and female reproductive tracts and appendixes.

Current knowledge of sex determination was built on gene by gene knock-in/knock-out in mice and transcriptomic analysis on pool of purified cells. The first method suffers from a very low throughput and produces a partial knowledge of the function of one or a limited set of genes. The second method caveat resides in the lack of specific reporters for the different cell types within the gonad and makes difficult the appreciation of the cell type heterogeneity, especially before sex determination. Moreover, a pool of purified cells results in an average message of non synchronous differentiating cells and thus blur the chronology of gene expressions.

In this study, we used transgenic mice carrying the Nr5a1-GFP reporter to isolate by FACS the gonadal somatic cells at five key stages of sex determination and gonad development in male. We proceeded to the RNA-sequencing of 391 single cells and identified the different cell types present in the developing testis and reconstructed their cell lineages. At E10.5, we detected one homogeneous progenitor cell population of $NR5A1^{+}$ cells expressing epithelial and stem cell marker genes, consistently with their bi-potential state before sex determination. From E11.5, one fraction of these cells activated a strong genetic program and initiated their differentiation as Sertoli cells and the other fraction of progenitor cells evolved transcriptionally and progressively start to express steroidogenic lineage markers such as *Arx* and *Pdgfra* and are restricted to the interstitial compartment of the testis. From E12.5, we detected few fetal Leydig cells differentiating from the interstitial progenitors.

This study represents the most granular transcriptomic database of gonadal somatic cells during early testis development. With these data, we contribute in the deeper understanding of the cell differentiation during sex determination.

How to analyse human or mouse Genome-scale CRISPR Knock-Out (GeCKO) datasets ?

Marc DELOGER[1,*], Pierre GESTRAUD[1,*], Raphaël MARGUERON[2] and Nicolas SERVANT[1]

[1] Institut Curie, PSL Research University, INSERM, U 900, MINES, ParisTech, F-75005, Paris, France

[2] Institut Curie, PSL Research University, CNRS, UMR 3215, INSERM, U 934, F-75005, Paris, France

[*] Equally contributed

Corresponding Author: marc.deloger@curie.fr

1 Introduction

Cas9 (Clustered Regularly Interspaced Short Palindromic Repeats (CRISPR) associated protein 9) target specificity is determined by a short 20bp sequence of a single guide RNA (sgRNA). Consequently, large scale oligo synthesis of guide sequences suggests a new way to interrogate gene function at a genome-wide scale. It has been shown that infecting cell with Cas9:sgRNAs libraries using lentiviruses can facilitate both positive and negative loss-of-function screening in mammalian cells [1, 2]. When Cas9-induced double-strand breaks are introduced into coding sequences, error-prone repair machinery will often introduce « indel » potentially leading to loss-of-function allele [3]. In order to check the Genome-scale CRISPR Knock-Out (GeCKO) screening modification efficiency, off-target modification rate, consistency between unique sgRNAs targeting the same genes and validation rate of screen hits, we decided to propose a full R package that goes from FASTQ files to an HTML report giving all quality control and statistical analysis results of your GeCKO experiments taking into account process data difficulties such as results ranking.

2 Data pre-processing, Quality Control and Statistical analysis

Original human GeCKOv2 library is composed of 123'411 sgRNAs (~ 6 per gene + 2000 negative controls). First, we propose to clean the library by merging redundant sequences and removing the ones that have multi-hits on hg38 genome, obtaining 113'761 guides. These sequences were detected (with or without mismatches) and counted on each sample sequenced reads (FASTQ files) in order to generate a count table as output for further analysis. Nine criteria are used to assess the experiment quality. An R markdown notebook is automatically generated with all the corresponding values/figures.

From the count table, data are normalised using classical RNAseq methods from R package edgeR. We propose to compute the normalisation factor either on the whole dataset or on the non-targeting guides only. Differential abundance of guides is estimated with the limma/voom framework. From the moderated t-statistics estimated by limma, we derived a one-sided p-value to test explicitly the depletion or the enrichment of guides. A gene-level score is then computed by the RRA method [4] and p-values are derived from a null distribution estimated from random genes sampled into non-targeting guides scores.

Acknowledgements

This project was done thanks to SiRIC-Curie funding and Raphaël MARGUERON's datasets

References

[1] Shalem O, Sanjana NE, Hartenian E, Shi X, Scott DA, Mikkelsen TS, Heckl D, Ebert BL, Root DE, Doench JG, Zhang F. Genome-scale CRISPR-Cas9 knockout screening in human cells. Science. 2014 Jan 3;343(6166):84-7.

[2] Sanjana NE, Shalem O, Zhang F. Improved vectors and genome-wide libraries for CRISPR screening. Nat Methods. 2014 Aug;11(8):783-4.

[3] Rodgers K, McVey M. Error-Prone Repair of DNA Double-Strand Breaks. J Cell Physiol. 2016 Jan;231(1):15-24.

[4] Kolde R, Laur S, Adler P, Vilo J: Robust rank aggregation for gene list integration and meta-analysis. Bioinformatics. 2012, 28: 573-580. 10.1093/bioinformatics/btr709

Functional characterization of human epidermis co-expression modules

Gwenaëlle LEMOINE[1], Marie Pier SCOTT-BOYER[1], Mickael LECLERQ[1], Arnaud DROIT[1]

[1] Computational Biology Laboratory - CHU de Québec - Université Laval, 2705 Boulevard Laurier, G1V 4G2, Québec, Canada

Corresponding Author: lemoine.gwenaelle@gmail.com

Underlying molecular mechanisms in human tissues tend to be more documented everyday for a certain number of them [1]. Usually studied in pathological context, deepening the understanding of molecular mechanisms lead to a better prevention and treatment strategies of related diseases. Still, many (patho)physiological reactions remain unpredictable [2,3] if solely observed at the molecular level.

However, the recent surge of high-throughput omics technologies (e.g. transcriptomics, proteomics, metabolomics) allowed the emergence of new discoveries by system-wide analyses. A popular approach to analyze these large data is the modeling of gene similarity based networks. Also called co-expression networks, they group highly dependent genes by modules. Those modules can mostly be related to and used to further understanding of phenotypic traits information. Besides summarizing gene expression, this method allow to functionally annotate unknown genes.

Here, we emphasis on epidermal data which remains one of the easiest tissues to explore because of its ease of access for sampling (unlike tissues from internal organs). Helped by the R package WGCNA, first step of the pipeline involves the network creation. It is based on previous defined similarity and a threshold consisting of numerical set value (called hard-thresholding), or a function with a defined parameter value (called soft-thresholding). Module detection inside the network is then performed in order to group highly co-expressed genes. This process is then repeated on several pertinent data sets in order to identify epidermis-specific modules and compare them against each other.

Because those modules are more likely to mediate phenotypic traits of interest [4], a further investigation aims at comparing the activity of epidermis-specific modules under different conditions. On one hand, skin-linked factors like UV exposure, skin diseases or scaring will be studied. On the other hand, we will look into more general element such as sex or age impact. Most data comes from available public databases, essentially GTEx [5] or GEO [6], and internal sequencing data of *in-vivo* skin samples will complete them. Finally, the main frame is to lead to a refining of module functional characterization, and therefore allow some supervised learning for ulterior prediction models.

Since transcriptomic impacts other omics stratum in many ways, the incoming challenge is the association of those outlined epidermis-specific gene modules with other biological networks such as protein-protein interactions networks or metabolic pathways in the same tissue [7].

References

[1] Glass et al. *Gene expression changes with age in skin, adipose tissue, blood and brain*. Genome Biology, 2013
[2] Sudharsana Sundarrajan, Mohanapriya Arumugam. *Weighted gene co-expression based biomarker discovery for psoriasis detection*. GeneVolume 593, Issue 1, 15 November 2016, Pages 225–234
[3] Kuehne et al. *An integrative metabolomics and transcriptomics study to identify metabolic alterations in aged skin of humans in vivo Andreas.* BMC Genomics, 2017
[4] Sipko van Dam, Urmo Võsa, Adriaan van der Graaf, Lude Franke, João Pedro de Magalhães. *Gene co-expression analysis for functional classification and gene–disease predictions*. Brief Bioinform, 2017
[5] The GTEx Consortium. *The Genotype-Tissue Expression (GTEx) project*. Nature Genetics 45, 580–585, 2013
[6] Edgar R, Domrachev M, Lash AE. *Gene Expression Omnibus: NCBI gene expression and hybridization array data repository*. Nucleic Acids Res., 207-10, 2002
[7] Serin Elise A. R., Nijveen Harm, Hilhorst Henk W. M., Ligterink Wilco. *Learning from Co-expression Networks: Possibilities and Challenges*. Frontiers in Plant Science, 2015

Integrated and flexible analysis of scRNA-seq data with Eoulsan scRNA-seq

Geoffray BRELURUT[1]*, Nathalie LEHMANN[1]*, Céline HERNANDEZ[1], Morgane THOMAS-CHOLLIER[1], Denis THIEFFRY[1], Stéphane LE CROM[2] and Laurent JOURDREN[2]

[1] École normale supérieure, PSL Research University, CNRS, Inserm, Institut de Biologie de l'École normale supérieure (IBENS), Equipe Biologie Computationnelle des Systèmes, 75005 Paris, France.
[2] École normale supérieure, PSL Research University, CNRS, Inserm, Institut de Biologie de l'École normale supérieure (IBENS), Plateforme Génomique, 75005 Paris, France.
* These authors contributed equally to this work

Corresponding Author: jourdren@biologie.ens.fr

Keywords: Single-Cell RNA sequencing, distributed computing, high-throughput sequencing, pipeline.

Summary

Single cell RNA sequencing (scRNA-seq) provides new opportunities to characterize cell transcriptomic heterogeneity, e.g. the discovery of new cell populations and the reconstruction of cell lineages. Sequence reads are produced from individual cells using two main approaches: full transcript amplification or partial amplification with single molecule tagging using Unique Molecular Identifiers (UMI) [1]. These different protocols combined with various computational analyses result into numerous possible computational workflows.

Here we present a new workflow dedicated to scRNA-seq data analysis. Still under development, this workflow is based on the *Eoulsan* framework [2], starts with raw FASTQ files, and encompasses various quality control, normalisation, read mapping, and more advanced steps.

The first steps of the analysis (from read filtering to expression estimation) can be distributed on computer clusters to greatly reduce processing time. Most of the popular frameworks or job schedulers like *Hadoop*, *TORQUE* or *HTCordor* are currently supported by *Eoulsan*. Noteworthy, we take full advantage of the *Docker-Galaxy* framework introduced in *Eoulsan 2* [3] to propose a series of modules allowing the analysis of both read-based and UMI-based data. Downstream steps include a wide range of tools for gene differential expression (*SCDE* [4], *SCDD* [5]), cell clustering and lineage reconstruction (*Monocle 2* [6], *Destiny 2* [7]).

The experimental design ofr an *Eoulsan scRNA-seq* workflow is stored in simple text file, while parameters are stored in a second xml file, ensuring flexibility and traceability. Keeping track of each successful step, this approach allows to swiftly resume large analyses upon trouble-shooting, and further ensure reproducibility.

In conclusion, *Eoulsan scRNA-seq* provides an integrated workflow for scRNA-seq data analysis on standalone workstations or on computer clusters. With its modular structure and distributed data processing, it can handle large amounts of data in a reproducible, yet flexible manner.

References

[1] Christoph Ziegenhain *et al.* Comparative Analysis of Single-Cell RNA Sequencing Methods. *Molecular Cell*, 65: 631-43, 2017.
[2] Laurent Jourdren *et al.* Eoulsan: A Cloud Computing-Based Framework Facilitating High Throughput Sequencing Analyses. *Bioinformatic,* 28:1542-3, 2012.
[3] http://www.outils.genomique.biologie.ens.fr/eoulsan2/
[4] Peter V. Kharchenko, Lev Silberstein, and David T. Scadden. Bayesian approach to single-cell differential expression analysis. *Nature Methods,* 11:740-2, 2014.
[5] Keegan D. Korthauer *et al.* A statistical approach for identifying differential distributions in single-cell RNA-seq experiments. *Genome Biology,* 17:222, 2016.
[6] Cole Trapnell *et al*. Pseudo-temporal ordering of individual cells reveals dynamics and regulators of cell fate decisions. *Nature Biotechnology,* 32:381-6, 2014.
[7] Philipp Angerer *et al.* Destiny: diffusion maps for large-scale single-cell data in R. *Bioinformatics*, 32: 1241-3, 2015.

Efficient data structure for indexing and similarity computation of nucleic sequences

Camille MARCHET[1], Arnaud MENG[2], Lolita LECOMPTE[1], AntoineLIMASSET[1], LucieBITTNER[2] and Pierre PETERLONGO[1]

[1] IRISA Inria Rennes Bretagne Atlantique, GenScale team
[2] Sorbonne Universités, UPMC Univ Paris 06, CNRS, Evolution Paris Seine – Institut de Biologie Paris Seine (EPS – IBPS), 75005 Paris, France

Corresponding author: camille.marchet@irisa.fr

1 Introduction

Sequencing of (meta-)genomes and (meta-)transcriptomes generates huge sets of sequences, that are often chopped in voluminous sets of k-mers for their further analysis. Such instances constitute challenges for high performance computation. To extract relevant pieces of information from the large data sets generated by current sequencing techniques, one must rely on extremely scalable methods and solutions. In this work we present a straightforward indexing structure called quasi-dictionary that scales to billions of elements and we propose direct applications based on k-mer diversity to explore (meta-)genomics/transcriptomics data sets.

2 Quasi-dictionary applications to sequencing data

The quasi-dictionary structure is tailored to handle huge quantities of k-mers to index. It relies on a MPHF library [1] combined with a fingerprint system. We present two applications of such a data structure for short reads data: SRC_COUNTER and SRC_LINKER. Both can be used to compare two samples or sample versus itself in particular when traditional methods do not scale. SRC_COUNTER links any read to its estimated abundance in a collection of samples. SRC_LINKER connects any read from a given sample to similar reads in the data set [2]. Furthermore we present an application of SRC_LINKER to long reads. Contrary to the previous applications, the input are long, erroneous reads from last generation of sequencing (PacBio, Nanopore).

3 Results

We present the quasi-dictionary data structure performances and its applications results and practical use cases, notably at work on marine data sets. We state the memory and time performances of the quasi-dictionary as well as the practical impact of its false positives. We also compare SRC_LINKER with state of the art tools for short and long reads for retrieving similarities between reads. As for short reads, we demonstrate our gain in scaling, that applies in particular in meta-genomics field. With long reads, we show the benefits of having a lightweight data structure to deal with high error rates with robustness, combined to scalability.

We highlight the fact that SRC_LINKER is one of the few tools that can be applied successfully to explore short as well as long reads sequencing data. Finally the data structure and its applications remain simple, which make them easily adaptable for further challenges.

4 Citations

References

[1] Antoine Limasset, Guillaume Rizk, Rayan Chikhi, and Pierre Peterlongo. Fast and scalable minimal perfect hashing for massive key sets. *arXiv preprint arXiv:1702.03154*, 2017.

[2] Veronika B Dubinkina, Dmitry S Ischenko, Vladimir I Ulyantsev, Alexander V Tyakht, and Dmitry G Alexeev. Assessment of k-mer spectrum applicability for metagenomic dissimilarity analysis. *BMC Bioinformatics*, 17(1):38, dec 2016.

Detection of poly-adenylation sites from RNA-Seq data

Cyril FOURNIER and Anamaria NECSULEA
Univ Lyon, Laboratoire de Biométrie et Biologie Évolutive, Université Claude Bernard Lyon1, 69622 Villeurbanne cedex, France

Corresponding author: cyril.fournier@univ-lyon1.fr

Submission

Transcriptome sequencing technologies (RNA-seq) are increasingly used to accurately and sensitively measure gene expression levels, across a wide variety of biological samples [1]. The utility of RNA-seq goes beyond the simple assessment of gene expression levels, as it allows detection and quantification of alternative transcript isoforms, as well as annotation of new genes and isoforms. Here, we propose a method to annotate cleavage and poly-adenylation sites of messenger RNAs and long non-coding RNAs [2,3] from RNA-Seq data.

Our computational pipeline comprises three steps:

- First, we efficiently search for polyA tails in RNA-Seq data, allowing for sequencing and biological errors. This step results in a complete set of reads with polyA tracts, which includes genuine 3' ends of transcripts but also encompasses polyA repeats present in the genome.
- Second, to filter out these repeats from the dataset and to enrich for genuine polyA tails, the reads are artificially cleaved to remove the polyA tail and the remainder is aligned on the reference genome. We then select the reads for which the polyA tract does not map on the genome, as expected given that polyA tails are added post-transcriptionally to messenger RNAs.
- Third, the predicted poly-adenylation and cleavage sites are assessed by checking the flanking nucleotides and the presence of a poly-adenylation signal, and finally compared to known cleavage sites.

The running time of this method with a sample of 100 millions reads of 100 bp is approximately of 20 minutes (test run on Ubuntu 16.04 LTS with 16Go RAM and i7-6700HQ CPU @ 2.60GHz). Thus, this method can be easily applied to the vast quantities of RNA-Seq data that are accumulating today. Our method is thus considerably faster than other programs that detect polyA sites from RNA-Seq data, such as *KLEAT* [4] and *ContextMap 2* [5].

We applied our method to a collection of RNA-Seq data, derived from four major organs (brain, kidney, liver and testis) and five developmental stages (from early organogenesis to adult and aged individuals), for mouse and rat. For a standard RNA-Seq sample, the fraction of reads with polyA tail detected is approximately 3/1000 reads. We annotated over 100,000 putative polyA sites in each species and we confirmed about 10,000 sites annotated in the Ensembl database and 3,000 sites using polyA-sequencing data, from *Derti et al* [6].

Ongoing analyses include comparative analyses of the predicted polyA sites across species, organs and developmental stages. We also plan to use our method to refine the 3' ends of long non-coding RNAs, whose annotation is currently challenging.

References

[1] Zhong Wang, Mark Gerstein, and Michael Snyder. RNA-Seq: a revolutionary tool for transcriptomics. *Nature reviews Genetics*, 10, jan 2009.

[2] Diana F. Colgan and James L. Manley. Mechanism and regulation of mRNA polyadenylation. *Genes & Development*, 11:2755–2766, 1997.

[3] Bin Tian and Joel H. Graber. Signals for pre-mRNA cleavage and polyadenylation. *Wiley Interdisciplinary Reviews: RNA*, 3(3):385–396, 2012.

[4] Birol I, Raymond A, Chiu R, Nip KM, Jackman SD, and Kreitzman M et al. Kleat: cleavage site analysis of transcriptomes. *Pac Symp Biocomput*, 20:347–358, 2015.

[5] Thomas Bonfert and Caroline C. Friedel. Prediction of poly(a) sites by poly(a) read mapping. *PLOS ONE*, 12(1):1–32, 01 2017.

[6] Adnan Derti, Philip Garrett-Engele, Kenzie D. MacIsaac, Richard C. Stevens, Shreedharan Sriram, Ronghua Chen, Carol A. Rohl, Jason M. Johnson, and Tomas Babak. A quantitative atlas of polyadenylation in five mammals. *Genome Research*, 22(6):1173–1183, 2012.

Développement d'une structure pour l'indexation et la compression de multi-génomes / New structure to index and compress multi-genomes

Clément AGRET[1,2], Manuel RUIZ[1] and Alban MANCHERON[2]

[1] UMR AGAP, CIRAD, Avenue Agropolis, 34398, Montpellier, FRANCE
[2] LIRMM, Campus St Priest, 161 Rue Ada, 34090 Montpellier, FRANCE

Corresponding Author: clement.agret@gmail.com

Alors qu'il a fallu treize ans et plusieurs millions de dollars pour séquencer et assembler les trois milliards de nucléotides composant le génome humain [1], un tel séquençage nécessite aujourd'hui seulement quelques jours et à peine quelques milliers de dollars. L'International Rice Research Institute (IRRI), dans le cadre du consortium GRISP (Global Rice Science Partnership), a initié un programme de séquençage de l'ensemble des variétés de riz, et aujourd'hui plus de 3000 génomes sont déjà disponibles [2].

Pour analyser et stocker efficacement cette importante masse de données, il est nécessaire de représenter les génomes dans une forme permettant leur consultation rapide, tout en économisant le plus possible l'espace nécessaire à leur stockage.

La structure de données qui fait actuellement le succès des méthodes d'analyse de génomes est le FM-index [3]. Il s'agit d'une structure compressée exploitant les propriétés de réorganisation des données de la Transformée de Burrows-Wheeler (BWT) [4] appliquées sur le génome à indexer. Cependant le FM-index n'est pas optimisé pour compresser une collection de génomes similaires. Ce qui veut dire que si l'on souhaite analyser les 3000 génomes de riz, on va devoir créer 3000 index pour pouvoir ensuite interroger chaque index l'un après l'autre.

Nous avons exploré et comparé les méthodes existantes (PanTools, TwoPaco, CHICO, etc.) permettant de construire efficacement un index commun aux 3000 génomes. Nous développons aussi une méthode basée sur un découpage des génomes par k-mers. Pour cela, nous avons dans un premier temps étudié l'évolution du nombre de k-mers communs entre différents génomes complets de riz. La représentation des ensembles de k-mers communs entre les différents génomes sous forme de diagrammes de Venn nous a confirmé la pertinence de notre approche. En raison du volume de données qui seront indexées, nous avons défini une représentation de l'ensemble des k-mers et de leurs présence/absence basée sur des structures succinctes. Les problématiques qui émergent sont d'une part la mise à jour dynamique de la structure, par exemple lorsqu'un génome est ajouté ou retiré de l'ensemble des génomes déjà indexés; et d'autre part la formulation et l'optimisation des requêtes que notre structure d'index doit permettre. Enfin, les problématiques liées à la représentation et l'exploration visuelle de l'index demeurent un sujet d'étude qu'il conviendra de traiter par la suite.

References

[1] International Human Genome Sequencing Consortium : Initial sequencing and analysis of the human genome. Nature, 409(6822):860–921, 2001.
[2] Li JY, Wang J, Zeigler RS: The 3,000 rice genomes project: new opportunities and challenges for future rice research. Gigascience 3, 8, 2014.
[3] Paolo FERRAGINA et Giovanni MANZINI: Opportunistic Data Structures with Applications. In Proceedings of the 41 st Annual Symposium on Foundations of Computer Science (FOCS), pages 390–398.
[4] Michael BURROWS et David WHEELER: A Block-Sorting Lossless Data Compression Algorithm. Rapport technique, 1994.

Epigenetic marks and the human transcriptome diversity

Guillaume DEVAILLY[1], Anna MANTSOKI[1] and Anagha JOSHI[1]
Developmental biology, The Roslin Institute, The University of Edinburgh, Easter Bush, Midlothian, EH25 9RG, Scotland, UK

Corresponding author: guillaume.devailly@roslin.ed.ac.uk

Links between epigenetic marks and transcription regulation have been increasingly explored in recent years. Major technical advancements allowed the generation of massive, under-exploited datasets by consortia and individual laboratories. We present a systematic integrative re-analysis of Roadmap Epigenomics RNA-seq, WGBS, DNAseI and histone marks ChIP-seq data in 33 human cell types. We combined two complementary approaches to link epigenetic marks at location of interests (transcription start sites, transcription end sites, exons) to transcription features (transcription level, exon inclusion ratio, cryptic transcription start sites): 1- for all genes/exons within a cell, and 2- for every single gene/exon across all 33 cell types.

Results generated include the confirmation of many known relationships (i.e. promoter DNA methylation is negatively correlated to gene expression, while H3K4me3 at promoter is positively regulated to gene expression) at a large scale, but also the confirmation or rejection of less accepted correlations. For example, while we did find a slight decrease of DNA methylation density at non-included region within a cell, we could not find any trend linking changes of exon methylation to changes of inclusion ratio when carrying the analysis across cells. A web portal will be developed in the near future to allow easy exploration of the results.

Acknowledgements

AJ is a Chancellor's fellow and AJ labs is supported by institute strategic funding from Biotechnology and Biological Sciences Research Council (BBSRC, BB/J004235/1). GD is funded by the People Programme (Marie Curie Actions FP7/2007-2013) under REA grant agreement No PCOFUND-GA-2012-600181.

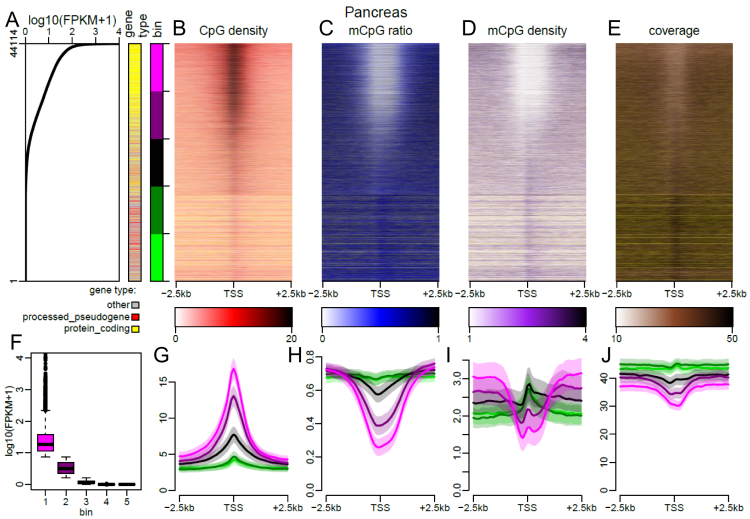


Fig. 1. Promoter DNA methylation and gene expression level in pancreas. **A.** Gene expression level in all 44114 genes annotated by Gencode. First side bar indicates the gene type (yellow, protein coding genes), the second side bar indicates the 5 bins used in panel F-J (purple: highly expressed genes, green: lowly expressed genes). Protein coding genes are globally more expressed than other types of genes. **B-E.** Stacked profiles of CpG density (B), the mCpG ratio (mCpG/CpG) (C), the mCpG density (D), and the WGBS coverage (E) at promoter, sorted according to their expression level. Promoter CpG density is positively correlated with expression level, mCpG ratio and mCpG density are negatively correlated to expression level. Coverage of CpG rich region is lower than at CpG poor regions. **F.** Boxplot of gene expression level in each of the 5 bins defined in A, and used in G-J. **G-J.** Average profile of CpG density (G), mCpG ratio (H), mCpG density (I) and WGBS coverage (J) +/- SEM for each bin of promoters.

Microbial diversity and plant cell wall-degrading enzyme dynamics during dew-retting of flax - one of the oldest applications of biotechnology to textile transformation

Christophe DJEMIEL[1], Sébastien GREC[1] and Simon HAWKINS[1]

[1] UGSF – Unité de Glycobiologie Structurale et Fonctionnelle, Université de Lille, 59000, Lille, France

Corresponding Author: christophe.djemiel@gmail.com

The existence of spun and colored flax fibers dating from the Upper Paleolithic suggests that man has long exploited the biotechnological natural process known as dew-retting to extract these fibers from the flax plant for textile production [1,2]. This process is achieved directly on the soil surface of the field [3,4]. Despite many studies of this process to evaluate the degree of retting [5], relatively little is known about (i) the composition and the evolution of the microflora population during retting, (ii) the kinetics of the microbial communities colonizing the plant material and (iii) the composition and the evolution of Carbohydrate Active enZymes degrading plant cell walls [6].

To improve our understanding of dew-retting, we first used a metabarcoding approach to identify the membership and structure of the microbial communities (focusing on bacteria and fungi). This approach also allowed us to identify (i) some potential bacterial major enzymatic functions related to carbohydrate degradation based on functional prediction using PICRUSt (http://picrust.github.io/picrust/) and (ii) a strong pattern of fungal trophic modes [7,8]. In a second step we developed a metatranscriptomic approach to access of the evolution of the exogenous (soil) and endogenous (plant) enzymatic arsenal potentially across the Tree of Life, involved in the degradation of carbohydrate and aromatic substances in decomposing plant matter.

The methodology used to explore microbial and enzymatic diversity using High Throughput Sequencing (Illumina system) will be described. We will then correlate colonization complexity dynamics to the progress in plant cell wall degradation. Finally, the microbial ecology of the retting process will be compared to other natural plant material (e.g. forest litter) degradation processes.

Acknowledgements

This work is funded within the framework of the collaborative French 'Future project' SINFONI.

References

[1] Roland, J.-C., Mosiniak, M. & Roland, D., 1995. Dynamique du positionnement de la cellulose dans les parois des fibres textiles du lin (Linum usitatissimum). Acta Botanica Gallica, 142(5), pp.463–484.
[2] Kvavadze, E. et al., 2009. 30,000-Year-Old Wild Flax Fibers. Science, 325(5946), pp.1359–1359.
[3] Md. Tahir, P. et al., 2011. Retting process of some bast plant fibres and its effect on fibre quality: A review. BioResources, 6(4), pp.5260–5281.
[4] Akin, D.E., 2013. Linen most useful: Perspectives on structure, chemistry, and enzymes for retting flax. ISRN Biotechnology, 2013, p.23.
[5] Martin, N. et al., 2013. Influence of the degree of retting of flax fibers on the tensile properties of single fibers and short fiber/polypropylene composites. Industrial Crops and Products, 49, pp.755–767.
[6] Lombard, V. et al., 2014. The carbohydrate-active enzymes database (CAZy) in 2013. Nucleic acids research, 42(Database issue), pp.D490-5.
[7] Langille, M.G. et al., 2013. Predictive functional profiling of microbial communities using 16S rRNA marker gene sequences. Nat Biotechnol, 31(9), pp.814–821.
[8] Nguyen, N.H. et al., 2016. FUNGuild: An open annotation tool for parsing fungal community datasets by ecological guild. Fungal Ecology, 20, pp.241–248.

Splicing Lore: Speeding up the identifications of splicing factors regulating alternative exons across physiological and pathological conditions.

HELENE POLVÈCHE[1], NAHED BOUCHOUICHA[1] and DIDIER AUBOEUF[1]

[1] LBMC, 46 Allée d'Italie, 69364 Lyon Cedex 07, France

Corresponding Author: helene.polveche@ens-lyon.fr

Alternative splicing (AS) leads to the production of different transcripts from each gene and relies on *splicing factors* (SFs) site recognition. Massive splicing variations are observed in many diseases. However, identifying which mechanisms that are responsible for these variations is challenging. To address this concern, we developed “Splicing Lore” whitch aims at predicting the SF(s) regulating the inclusion rate of exons from a selected list of alternative exons.

Splicing Lore contains the list of exons regulated by 64 SFs, which corresponds to 94 publicly datasets. These datasets were analyzed with an homemade pipeline (FaRLine) allowing to quantify the effect of SF downregulation on the inclusion rate of human exons. The results were stored in a MySqL *database*, Splicing Lore DB that will be available for download.

An user-friendly *interface* allows to enter a list of exons and interrogate Splicing Lore DB. This allows users to know which SF(s) can regulate at least part of the exons of interest. Splicing Lore provides support toward the identification of the molecular mechanisms driving splicing variations across physiological or pathological situations.

Keyword : *alternative splicing, splicing factor, RNAseq, database, interface*

A method for DNA virus detection and quantification during pregnancy based on noninvasive prenatal testing whole genome sequencing results

Virginie CHESNAIS[1], Alban OTT[1], Emmanuel CHAPLAIS[1], Samuel GABILLARD[1], Christelle VAULOUP-FELLOUS[2], Alexandra BENACHI[3], Jean-Marc COSTA[4], and Eric GINOUX[1]

[1] Life&Soft, 8b avenue Descartes, 92350, Plessis-Robinson, France
[2] Hôpital Paul Brousse, 12 avenue Paul Vaillant Couturier, 94800, Villejuif, France
[3] Hôpital Antoine-Béclère, 157 rue de la Porte de Trivaux, 92140, Clamart, France
[4] Laboratoire CERBA, 11 rue de l'Équerre, 95310, Saint-Ouen-l'Aumône, France

Corresponding Author: vchesnais@lifeandsoft.com

Human cytomegalovirus (HCMV) is a DNA β-herpevirus of critical importance to human health during pregnancy. HCMV primary infection of pregnant women could lead to congenital infection of fetus and could have severe clinical complication in child [1]. With 10–100 billion fragments per milliliter of plasma, circulating cell-free DNA is an information-rich window into human physiology, with rapidly expanding applications in genetic prenatal diagnosis. The whole genome sequencing (WGS) of cell-free plasma DNA is classically used to diagnose fetal aneuploidy during pregnancy. Because WGS has also become a standard tool for pathogen discovery in biological samples [2], the purpose of this study is to propose a new method to detect and quantify circulating viral DNA during pregnancy using the same sequencing results as noninvasive prenatal testing whole genome sequencing data.

Our approach used the human unmappable reads to search viral specific reads and quantify the species viral load on low depth sequencing data, by combining multiple steps of alignment on reference genome to filter host reads and mapped exogenous reads on in-house validated reference genome of different viral species. Different quality steps checked that sequencing results had a sufficient quality to be analyzed and that the alignment on targeted viral genome is homogenous. We used sequenced ranged samples to calibrate our method and determine a mean depth threshold to classify infected and non-infected samples with a sensitivity of 100% [88%-100] and a specificity of 100% [94%-100%]. Because the mean depth was linearly correlated to the theoretical viral concentration of each sample, we constructed a model able to determine the viral load of each sample.

Our method was then applied to a cohort of 538 pregnant women to validate our approach with real clinical samples. In this validation cohort, we found two positive samples for HCMV with very low viral load. The serologic status of these two samples revealed that the patients were immunized against HCMV at the beginning of their pregnancy, suggesting a possible viral reactivation or secondary infection. In the same cohort, we also found three samples positive for HBV. For two of them the serologic status was known: these women were positive for HBV infection, confirming that our pipeline can detect several targeted viral species.

In summary, this study demonstrates the whole potential benefit of WGS based monitoring for pregnant women to perform complete prenatal diagnosis based on a single test.

References

[1] W. Britt, Manifestations of human cytomegalovirus infection: proposed mechanisms of acute and chronic disease, Curr. Top. Microbiol. Immunol. 325 (2008) 417–470.
[2] Q. Wang, P. Jia, Z. Zhao, VirusFinder: software for efficient and accurate detection of viruses and their integration sites in host genomes through next generation sequencing data, PloS One. 8 (2013) e64465. doi:10.1371/journal.pone.0064465.

RINspector: a Cytoscape app that combines centrality analyses with DynaMine flexibility prediction

Guillaume Brysbaert[1], Kevin Lorgouilloux[1], Wim Vranken[2] and Marc F Lensink[1]

[1] University of Lille, CNRS UMR8576 UGSF, F-59000 Lille, France
[2] Interuniversity Institute of Bioinformatics in Brussels (IB)[2], ULB-VUB and Structural Biology Brussels, VUB, Brussels, Belgium

Corresponding Author: guillaume.brysbaert@univ-lille1.fr

Deciphering the function of a protein is an intricate task that needs integrated approaches. In silico methods based on sequence and structure give good insights to solve this problem but few permit to integrate their results. An approach based on residue interaction networks analyses and in particular centrality analyses showed to give clues to the involvement of key residues in function and folding of a protein - e.g. [1, 2]. These networks are built from a protein structure where nodes are residues and edges are detected interactions between residues. Next to these methods, flexibility of a backbone and changes upon mutation provide additional information that help to select such key residues in the design of mutagenesis experiments and to unravel the function of a protein [3, 4].

We developed an app called RINspector for the Cytoscape network analysis and visualization software [5] to analyze residue interaction networks and visualize flexibility predictions associated to a protein chain. This app performs centrality analyses based on shortest path lengths, with associated Z-scores, and queries the DynaMine server [6, 7] to retrieve a prediction of the dynamic of a protein chain. Results can be visualized on an interactive flexibility graph which permits selection of residues that can be mutated to compare the new graph with the wild type. A connection is made between the flexibility graph, the residue interaction network and the protein structure (if structureViz app/Chimera installed [8, 9]) to select a residue simultaneously in the three representations thus to directly visualize this residue in its context. Our tool can help with the rapid identification of key residues for protein function and stability.

RINspector is available in the Cytoscape app store.

References

[1] del Sol, A., et al., *Residues crucial for maintaining short paths in network communication mediate signaling in proteins.* Mol Syst Biol, 2006. **2**: p. 2006 0019.
[2] Amitai, G., et al., *Network analysis of protein structures identifies functional residues.* J Mol Biol, 2004. **344**(4): p. 1135-46.
[3] Teague, S.J., *Implications of protein flexibility for drug discovery.* Nat Rev Drug Discov, 2003. **2**(7): p. 527-41.
[4] Golovanov, A.P., et al., *Structural consequences of site-directed mutagenesis in flexible protein domains: NMR characterization of the L(55,56)S mutant of RhoGDI.* Eur J Biochem, 2001. **268**(8): p. 2253-60.
[5] Shannon, P., et al., *Cytoscape: a software environment for integrated models of biomolecular interaction networks.* Genome Res, 2003. **13**(11): p. 2498-504.
[6] Cilia, E., et al., *From protein sequence to dynamics and disorder with DynaMine.* Nat Commun, 2013. **4**: p. 2741.
[7] Cilia, E., et al., *The DynaMine webserver: predicting protein dynamics from sequence.* Nucleic Acids Res, 2014. **42**(Web Server issue): p. W264-70.
[8] Morris, J.H., et al., *structureViz: linking Cytoscape and UCSF Chimera.* Bioinformatics, 2007. **23**(17): p. 2345-7.
[9] Pettersen, E.F., et al., *UCSF Chimera--a visualization system for exploratory research and analysis.* J Comput Chem, 2004. **25**(13): p. 1605-12.

Automated generation and analysis of parametric kinetic models obtained from biochemical interaction maps

Marion BUFFARD[1, 2], Oscar O. ORTEGA[3], Carlos F. LOPEZ[3] and Ovidiu RADULESCU[1]

[1] DIMNP, 163 rue Auguste Broussonnet, 34090, Montpellier, France

[2] Master of Sciences and Digital Technologies for Healthcare BioInformatics, Montpellier University, 34090, Montpellier, France

[3] Center for Quantitative Sciences, Vanderbilt University, 2220 Pierce Ave, Tennessee 37232, Nashville, United States

Corresponding Author: marion.buffard@etu.umontpellier.fr

The biochemical interaction map database, KEGG [1], contains approximately 500 Pathway maps. Creating parametric kinetic models manually for all of the interactions contained within all of the available maps would be tedious and would require the knowledge of tens of thousands of unknown kinetic parameters. A few approaches have been proposed to perform this task automatically such as for instance Path2Models [2]. However, these approaches do not cope with parametric problem. We propose a new method to automate this process including the generation of semi-quantitative kinetic laws. The analysis of such large dynamic models could allow for increased comprehension of the different phenotype or behavior of the cell and could elucidate new key proteins.

Using PySB [3], a framework for building mathematical models of biochemical pathways, we develop a tool that generates biochemical kinetic models from static interaction maps. The models are built as Python programs using PySB libraries for mechanistic interactions. PySB allows one to divide models into modules and to call libraries of reusable elements that encode standard biochemical actions. Species in the PySB model can be described with multiple sites and states, which reduce considerably the numbers of entities to declare.

To each interaction we associate a semi-quantitative parameter (integer) representing the order of magnitude of the interaction timescale. The values of these integers rank the interactions according to their timescales. Each interaction of the static model provides a system of rules. These rules are parametrized consistently with the interaction timescale orders. We consider that all internal processes, with exception to the limiting step, are more rapid than the timescale of the interaction. The choice of the limiting step can correspond either to quasi-equilibrium or to quasi-stationary conditions. The parametrized kinetic models are further analyzed. We are interested in classifying possible states and transitions between these states. We use tropical equilibration branches, a concept recently introduced in [4], as proxies for metastable states of the biochemical system.

As case studies we consider signaling models MAPK and TRAIL to compare our model to known models created using the classical approach.

Acknowledgements

This participation in the 18th JOBIM has been funded with thanks to Labex Numev (http://www.lirmm.fr/numev/).

References

[1] Hiroyuki Ogata, Susumu Goto, Kazushige Sato, *et al.* KEGG: Kyoto encyclopedia of genes and genomes. *Nucleic acids research*, vol. 27, no 1, p. 29-34, 1999.

[2] Finja Büchel, Nicolas Rodriguez, Nei Swainstonl, *et al.* Path2Models: large-scale generation of computational models from biochemical pathway maps. *BMC systems biology*, vol. 7, no 1, p. 116, 2013.

[3] Carlos F. Lopez, Jeremy L. Muhlich, John A. Bachman, *et al.* Programming biological models in Python using PySB. *Molecular systems biology*, vol. 9, no 1, p. 646, 2013.

[4] Satya Swarup Samal, Aurélien Naldi, Dima Grigoriev, *et al.* Geometric analysis of pathways dynamics: application to versatility of TGF-β receptors. *Biosystems*, vol. 149, p. 3-14, 2016.

Étude de processus de coalescence dans les paysages contemporains : bibliothèques template C++ pour le Calcul Bayésien Approché.

Arnaud BECHELER[1], Camille CORON[2] et Stéphane DUPAS[1]

[1] EGCE, Bâtiment 13, Campus CNRS, 1 Avenue de la Terrasse, 91198 Gif-sur-Yvette, France
[2] LMO, Faculté des Sciences d'Orsay Université Paris-Sud, F-91405 Orsay Cedex, France

Correspondant : Arnaud.Becheler@egce.cnrs-gif.fr

1 Introduction

Les modalités d'évolution récente des populations contemporaines deviennent une question majeure à l'heure de quantifier les impacts des changements globaux. Par exemple, dans le cadre d'une espèce invasive comme le frelon à pattes jaunes (*Vespa velutina*) en France, l'analyse d'un jeu de données génétique peut permettre la compréhension des réactions de l'espèce à son nouvel environnement.

En couplant des modèles démographiques spatialement explicites à des modèles de coalescence, il est possible d'inférer les paramètres des lois de croissance et de dispersion de l'espèce par Calcul Bayésien Approché (*Approximate Bayesian Computation*, ABC) [1]. Des modèles environnementaux permettent de représenter les lois de croissances ou de dispersion comme une fonction des patrons paysagers locaux [2].

Toutefois, les méthodes de simulation antérieures reconstruisent le coalescent jusqu'à l'ancêtre commun le plus récent (MRCA) des gènes échantillonnés [3]. Or, le MRCA étant situé dans une fenêtre spatio-temporelle lointaine, cela force à renseigner des processus historiques anciens peu informés et/ou peu informatifs. La nouvelle méthode présentée ici permet d'éviter les coûts d'une prise en compte de l'histoire ancienne (coût en hypothèses, en calcul et en données) en recentrant l'analyse sur les processus de coalescence très récents, c'est à dire ceux qui permettront d'informer l'histoire invasive du frelon asiatique.

2 Méthode

La nécessité d'envisager de nombreux modèles de différents niveaux de complexité, conjointement au besoin de performance imposé par l'ABC, nous a mené à développer des bibliothèques template C++ pour la simulation de processus de coalescence. Ces bibliothèques génériques, modulaires et extensibles, sont destinées à être rendues libres et offrent dans la définition du modèle simulatoire une liberté et une efficacité à notre connaissance sans équivalents actuels.

Lors d'une simulation, la coalescence peut être interrompue brutalement à la date où les données deviennent trop insuffisantes pour renseigner le processus, menant ainsi à une forêt de généalogies de gènes partitionnant le jeu de données génétique. Forêt de coalescents simulée et jeu de données observé sont alors convertis en partitions floues sans perte d'information, suite à quoi la procédure ABC permet d'accepter les valeurs de paramètres ayant servi à la simulation si la distance de transfert floue [4] calculée entre la partition simulée et la partition observée est inférieure à un certain seuil.

Références

[1] Estoup, Arnaud, et al. Combining genetic, historical and geographical data to reconstruct the dynamics of bioinvasions: application to the cane toad Bufo marinus. *Molecular ecology resources* (10.5):886-901, 2010.
[2] He, Qixin, Danielle L. Edwards, and L. Lacey Knowles. Integrative testing of how environments from the past to the present shape genetic structure across landscapes. *Evolution* (67.12): 3386-3402, 2013.
[3] Currat, Mathias, Nicolas Ray, and Laurent Excoffier. SPLATCHE: a program to simulate genetic diversity taking into account environmental heterogeneity. *Molecular Ecology Notes* (4.1): 139-142, 2004.
[4] Campello, Ricardo JGB. Generalized external indexes for comparing data partitions with overlapping categories. *Pattern Recognition Letters* (31.9): 966-975, 2010.

PROqPCR : a Shiny web application for PROcessing of qRT-PCR data

Mathilde SAUTREUIL[1] and Caroline BÉRARD[2]
[1] LMRS UMR 6085, Normandie Université, Avenue de l'Université, BP.12, F76801 Saint-Étienne-du-Rouvray, France
[2] LITIS EA 4108, Normandie Université, Avenue de l'Université, F76801 Saint-Étienne-du-Rouvray, France

Corresponding author: mathilde.sautreuil@etu.univ-rouen.fr

Summary : On this poster, we present PROqPCR, a user friendly tool enabling the PROcessing of quantitative Real-Time Polymerase Chain Reaction (qRT-PCR) data. The qRT-PCR is a common-used technique, which allows to quantify the transcriptome for pre-defined specific genes [1]. Despite the emergence of next generation sequencing, the qRT-PCR remains widely used in the laboratories. Indeed, one use of this experiment is, for instance, the mandatory validation of the results when high-throughput experiment is performed.

When a qRT-PCR experiment is performed, many files are created by the quantitative PCR instrument, and there is a long and repetitive treatment to do to carry out the analysis (averaging over replicates, normalization). That's why we automated it in a web application. This application allows biologists to perform easily the processing of qRT-PCR data by providing only the experimental design and the files created by the software of the quantitative PCR instrument.

At the end of the analysis, the biologist can visualize and explore the results from several graphs. Three different graphs are proposed and can be downloaded in png format:

1. Barplot of all conditions for a given gene;
2. Barplot of all genes for a given condition;
3. Comparison of barplots for selected genes in all conditions.

Moreover, two graphs are also provided in order to facilitate the comparison between the results of qRT-PCR with that of a RNA-seq experiment.

PROqPCR is written in R using the Shiny library, which provides access to powerful R-based functions and libraries through a simple user interface.
PROqPCR is free, open source and is available at https://qpcrapp.shinyapps.io/proqpcr/.

References

[1] C. A. Heid, J. Stevens, K. J. Livak, and P. M. Williams. Real time quantitative PCR. *Genome Research*, 6(10):986–994, January 1996.

Bivariate Negative Binomial Mixture Model for the analysis of RNA-seq data

Mathilde SAUTREUIL[1], Nicolas VERGNE[1], Antoine CHANNAROND[1], Angelina ROCHE[3], Gaëlle CHAGNY[1] and Caroline BÉRARD[2]

[1] LMRS UMR 6085, Normandie Université, Avenue de l'Université, BP.12, F76801 Saint-Étienne-du-Rouvray, France

[2] LITIS EA 4108, Normandie Université, Avenue de l'Université, BP.12, F76801 Saint-Étienne-du-Rouvray, France

[3] CEREMADE UMR CNRS 7534, Université Paris Dauphine, Place du Maréchal de Lattre de Tassigny, 75016 Paris, France

Corresponding author: mathilde.sautreuil@etu.univ-rouen.fr

Summary : RNA sequencing constitutes a method of choice to quantify gene expression. This experiment provides discrete read counts assigned to target genome regions measuring the expression level. Our goal is to perform differential analysis from these data, that is to say compare the counts of a given region between two conditions. Many methods of differential analysis were developed to compare two conditions using mainly statistical tests [1,2]. Here we propose a mixture model to search the differentially expressed genes. The idea is to classify the genes into three groups : one group where expression is the same between the two conditions and two groups where expression is higher in a condition than in the other one.

We developed a Bivariate Negative Binomial Mixture Model. Let $X_t = (X_{1t}, X_{2t})$ be the counts for each gene *t* in the conditions one and two, and Z_t the group this gene belongs. As X_t is bivariate, we have proposed to work with a mixture model in two dimensions. Moreover, the RNA-seq data are modeled using a Negative Binomial distribution because these data are discrete and overdispersed.

As written in Shi and Valdez [3], $X = (X_1, X_2)$ follows a Bivariate Negative Binomial distribution if there exists the independent variables Y_1, Y_2, Y_3 such as $Y_i \sim \mathcal{NB}(\theta_i, \alpha_i)$ with $i = 1, 2, 3$ and

$$\begin{cases} X_1 = Y_1 + Y_3 \\ X_2 = Y_2 + Y_3. \end{cases}$$

Thus, $X_i \sim \mathcal{NB}(\theta_i + \theta_3, \alpha_i + \alpha_3)$, with $\theta_i = \frac{\theta_3 \alpha_i}{\alpha_3}$.

The Bivariate Negative Binomial Mixture Model is written as :

$$\mathbb{P}(x|\psi) = \sum_{k=1}^{K} \pi_k MNB(x|(\theta_k, \alpha_k)), \text{ with}$$

- K the number of groups (here $K = 3$ as written before)
- MNB is the Multivariate Negative Binomial distribution
- $\psi = (\pi_1, ..., \pi_K, \theta_1, ..., \theta_K, \alpha_1, ..., \alpha_K)$, where $\theta_k = (\theta_{k1}, \theta_{k2}, \theta_{k3})$ et $\alpha_k = (\alpha_{k1}, \alpha_{k2}, \alpha_{k3})$
- $0 < \pi_k \leq 1, k = 1, ..., K$ et $\sum_k \pi_k = 1$.

The parameters are estimated with the EM algorithm [4]: π and θ have explicit estimators but α is estimated numerically. Then the genes are classified in one of the three groups according to the Maximum A Posteriori rule based on the posterior probabilities.

Through a simulation study, we will compare the obtained results with a Bivariate Gaussian Mixture Model, a Bivariate Poisson Mixture Model and the model we have developed. Moreover, we will compare our model with methods commonly used like DESeq2 [1]. Finally, we will apply the model on real data from an RNA-seq experiment on the *Phaeodactylum Tricornutum* diatom. In order to handle with replicates, we propose to average the replicates which is rounded at the closer integer.

References

[1] Michael I. Love, Wolfgang Huber, and Simon Anders. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biology*, 15:550, 2014.

[2] Paul L. Auer and Rebecca W Doerge. A Two-Stage Poisson Model for Testing RNA-Seq Data. *Statistical Applications in Genetics and Molecular Biology*, 10(1), 2011.

[3] Peng Shi and Emiliano A. Valdez. Multivariate Negative Binomial Models for Insurance Claim Counts. SSRN Scholarly Paper ID 2175226, Social Science Research Network, Rochester, NY, November 2012.

[4] A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society, Series B*, 39(1):1–38, 1977.

Quality evaluation of the mapping of small RNA-footprinting reads

Pauline Fourgoux[1], Etienne Delannoy[1], Claire Lurin[1], Guillem Rigaill[1], Véronique Brunaud[1]

[1] Institute of Plant Science Paris-Saclay (IPS2), Bât. 630 Plateau du Moulon rue Noetzlin, 91192, Gif sur Yvette cedex, FRANCE

Corresponding Author: pauline.fourgoux@u-psud.fr

Pentatricopeptide repeat (PPR) proteins are modular proteins which bind RNA [1]. They are found in every eukaryotes with a high number in Plants. There are about 500 PPRs in Arabidopsis. PPRs are key regulators in mRNA processing of organelles. For example they are involved in splicing, editing and stability[1]. The goal of our project is to detect RNA binding sites of PPRs using RNA-footprinting experiments. Arabidopsis thaliana will be the plant model for this study. PPR binding sites are small sequences of 8-30bp. Thus reads obtained from RNA-footprinting experiments are small. Correctly mapping those reads on chloroplast and mitochondrial genomes is difficult [2] yet crucial for our project.

Indeed it is important to correctly map those small reads to identify real PPR binding sites. Mapping small RNAs can be complicated, because they can match at different positions on the genome. Default parameters of mapping tools do not allow multiple hits and they can not be used in this case. In the case of PPRs, there are different causes to explain multiple hits. Firstly, some regions are repeated regions. Such regions lead to artefactual multiple hits. Secondly, some PPR proteins can bind to several identical RNA [4] at different position on the genome. In that case, we get “real” multiple hits.

The goal of our study is to identify the best tool to map our small RNA reads. We used pIRS [3] to simulate data and assess the mapping quality. pIRS allows to include SNP and InDel error rate parameters. Those parameters are important, because the PPR recognition code is inaccurate [1]. The analysis of RNA-Seq simulation showed two results. Firstly, they enabled to identify complex regions prone to multiple hits. We excluded those regions from the rest of our analysis. Secondly, they allowed us to calibrate quality threshold to use in case of multiple hits.

To conclude this precise quality assessment of mapping tools for small reads will facilitate the identification of real PPR binding sites using peak detection approach mostly developed for ChIP-Seq. Finding those sites is the main goal of our project. In particular, some of these sites can be related to cytoplasmic male sterility. Finding these sites will provide easy ways of producing F1 hybrids which are of high agronomical value.

References

[1]Manna, S. (2015). An overview of pentatricopeptide repeat proteins and their applications. *Biochimie*, *113*, 93-99.

[2]Johnson, N. R., Yeoh, J. M., Coruh, C., & Axtell, M. J. (2016). Improved placement of multi-mapping small RNAs. *G3: Genes| Genomes| Genetics*, *6*(7), 2103-2111.

[3]Hu, X., Yuan, J., Shi, Y., Lu, J., Liu, B., Li, Z., ... & Yue, Z. (2012). pIRS: Profile-based Illumina pair-end reads simulator. *Bioinformatics*, *28*(11), 1533-1535.

[4]Barkan, A., Rojas, M., Fujii, S., Yap, A., Chong, Y. S., Bond, C. S., & Small, I. (2012). A combinatorial amino acid code for RNA recognition by pentatricopeptide repeat proteins. *PLoS Genet*, *8*(8), e1002910.

Medical diagnosis pipelines on the new AP-HP bioinformatics platform

Jocelyn BRAYET[1], Camille BARETTE[1], Mathieu BARTHELEMY[1], Vivien DESHAIES[1] and Alban LERMINE[1]

[1] Bioinformatic platform of AP-HP, 5 rue Santerre (Hôpital Rothschild), 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 million patients.

One year ago, a new bioinformatics platform was created for multiple missions: the progressive centralization of the storage of genomic data produced by hospitals, their analyses in controlled and standardized workflows and the provisioning of tools for results exploitation.

In this abstract, we present our solution to develop and standardize medical diagnosis pipelines. The goal of this project is to propose a development environment for tools and parameters testing and an analysis environment with fixed pipelines and for results reproducibility. We decided to use *Docker* [1], *Galaxy* [2], and *Snakemake* [3] technologies. Docker allows to eliminate tool dependencies problems and sets a version tool in an image. We created a Docker image per tool per version. Galaxy is a web-based platform, designed to provide easy access to a versatile toolbox for biological users. Using Galaxy interface, scientists can test tools, tool versions and parameters but also design new workflows. When the pipeline is validated, Snakemake rules are written, versioned and tagged. Snakemake is a workflow management system with implicit rule implementation (input and output logic). The advantage over Make is the capability to allow Python to be interspersed through the pipeline and thereby reducing a lot of ambiguity [4]. As other pipeline frameworks, error recovery, automatic parallelization and workflow integrity features are included in Snakemake. With this project, on one hand, those Docker images are executed by a Galaxy instance. On the other hand, those same Docker images are run by Snakemake pipelines. To ease medical diagnosis routine, AP-HP's scientists are able to execute tagged workflows with their data through a web interface.

Furthermore, every Docker tools and Snakemake pipelines are tested with *GitLab-CI* [5]. GitLab is a web-based repository manager. GitLab-CI provides continuous integration features, so that when a new tool version is added in a Docker image, GitLab will build, run the container and automatically test the tool.

Key words: *Galaxy, Docker, Snakemake, diagnosis pipelines and continuous integration*

References

[1] Web site: https://www.docker.com/
[2] Enis A, Dannon B, Marius VDB, Daniel B, Dave B, Martin C, John C, Dave C, Nate C, Carl E, Björn G, Aysam G, Jennifer HJ, Greg VK, Eric R, Nicola S, Nitesh T, James T, Anton N, and Jeremy G. The Galaxy platform for accessible, reproducible and collaborative biomedical analyses: 2016 update. *Nucleic Acids Research* (2016) 44(W1): W3-W10 doi:10.1093/nar/gkw343
[3] Johannes Köster, Sven Rahmann; Snakemake—a scalable bioinformatics workflow engine. *Bioinformatics* 2012; 28 (19): 2520-2522. doi: 10.1093/bioinformatics/bts480
[4] Jeremy Leipzig; A review of bioinformatic pipeline frameworks. *Brief Bioinform* 2017; 18 (3): 530-536. doi: 10.1093/bib/bbw020
[5] Web site: https://about.gitlab.com/gitlab-ci/

Mathematical modeling of a genetic network controlling the regulation of Fe-S biogenesis

Firas HAMMAMI[1,2], Frédéric BARRAS[1], Pierre MANDIN[1] and Elisabeth REMY[2]

[1] Laboratoire de Chimie Bactérienne, Aix Marseille Université–CNRS, UMR 7283, Institut de Microbiologie de la Méditerranée, CNRS, 31 Chemin Joseph Aiguier, 13009 Marseille, France

[2] Institut de Mathématiques de Marseille, Aix-Marseille Université, CNRS, Centrale Marseille, I2M UMR 7373, Marseille, France.

Corresponding Authors: elisabeth.remy@univ-amu.fr or pmandin@imm.cnrs.fr

Iron-sulfur (Fe-S) clusters are cofactors conserved in all domains of life. A-Type Carriers (ATC) proteins play a crucial role in Fe-S biogenesis by delivering clusters to their targets [1]. In *E. coli*, the ErpA ATC protein is essential in aerobic growth, whereas NfuA plays an important role in stress conditions, such as oxidative stress or iron deficiency.

Recent work in our laboratory showed that *erpA* expression is regulated by IscR, the main Fe-S cluster homeostasis regulator, and by the non-coding RNA RyhB, expressed in iron deficient conditions [2]. While both regulators repress *erpA and nfuA* expression in opposite conditions in regards to iron concentration, these genes present different expression profiles.

In order to understand the mechanisms underlying *erpA* and *nfuA* regulation, we used a mathematical modeling approach relying on logical formalism. This model takes into account environmental conditions that perturb Fe-S biogenesis such as iron and oxygen levels. The model was validated with experimental data, allowing us to make predictions in different mutants and to infer qualitative network properties. For instance, the model suggests different inhibition thresholds for RyhB.

(Work in progress)

References

[1] Roche B, Aussel L, Ezraty B, Mandin P, Py B, Barras F. Iron/sulfur proteins biogenesis in prokaryotes: formation, regulation and diversity. *Biochim Biophys Acta*, Mar;1827(3):455-69, 2013

[2] Mandin P, Chareyre S and Barras F. A Regulatory Circuit Composed of a Transcription Factor, IscR, and a Regulatory RNA, RyhB, Controls Fe-S Cluster Delivery. *Mbio*, Sep 20;7(5), 2016

First gene-annotation enrichment analysis based on bacterial core-genome variants: Insights into mammalian and avian host adaptation of *Salmonella* serovars

Meryl VILA NOVA[1*], Kevin DURIMEL[1*], Arnaud FELTEN[1], Laurent GUILLER[1], Michel-Yves MISTOU[1] and Nicolas RADOMSKI[1]

[1] Laboratory for food safety (Anses), 14 rue Pierre et Marie Curie, 94700, Maisons-Alfort, France
*Poster presenters

Corresponding Author: nicolas.radomski@anses.fr

1 Introduction

Most of bacterial genomic studies exploring the host adaptation focus on the accessory genome describing how gain and loss of genes explain evolution processes leading colonization of new habitats. In this context, we propose a robust phylogenetic inference based on single nucleotide polymorphisms (SNPs) and recombination events, identification of fixed SNPs and small insertions/deletions (InDels) distinguishing homoplastic and non-homoplastic core-genome variants, and gene-annotation enrichment analyses in order to describe the main metabolism pathways impacted by these fixed variants during adaptation of *Salmonella* enterica subsp. Enterica to multi-host (*S.* Enteritidis), mammalian (*S.* Dublin), and avian (*S.* Pullorum and *S.* Gallinarum) hosts [1].

2 Results

The developed workflow 'VARCall-GO' produced a robust phylogenetic inference based on SNPs. The monophyletic clade S. Dublin diverged to the polyphyletic clade *S.* Enteritidis which includes the divergent clades *S.* Pullorum and *S.* Gallinarum. Firstly, this workflow gave the opportunity to detect intragenic and non-homoplastic fixed variants supporting the phylogenetic reconstruction [2]. Secondly, it identified representative metabolic pathways related to the host adaptation using the first gene-annotation enrichment analysis [3] based on bacterial core-genome variants. The host adaptation of *Salmonella* serovars were driven by fixed variants impacting mainly biosynthetic and metabolic processes of carbon sources alternative to glycose, amino acids, and ion transport, especially potassium.

3 Conclusion

We propose a new core-genome approach coupling identification of fixed SNPs and InDels with regards to the inferred phylogenetic clades, and gene-annotation enrichment analysis in order to describe the adaptation of *Salmonella* serovars Dublin, Enteritidis, Pullorum, and Gallinarum. All these generic bioinformatic tools can be applied on other bacterial genera without additional developments.

Acknowledgements

This work was supported by the COMPARE European project aiming to provide genomic tools in the context of outbreak detection among humans and animals worldwide. *Horizon 2020 research and innovation programme* under grant agreement No. 643476

References

[1] Langridge GC, Fookes M, Connor TR, Feltwell T, Feasey N, Parsons BN, et al. Patterns of genome evolution that have accompanied host adaptation in *Salmonella*. Proc. Natl. Acad. Sci. 2015;112:863–8.
[2] Aberer AJ, Pattengale ND, Stamatakis A. Parallelized phylogenetic post-analysis on multi-core architectures. J. Comput. Sci. 2010;1:107–14.
[3] Rachael P Huntley, Tony Sawford, Maria J Martin, Claire O'Donovan; Understanding how and why the Gene Ontology and its annotations evolve: the GO within UniProt. *Gigascience* 2014; 3 (1): 1-9. doi: 10.1186/2047-217X-3-4

A phylogenomic network to decipher bacterial adaptation through horizontal gene transfer

Damien Richard[1,2,3], Virginie Ravigné[1], Aude Chabirand[2], Olivier Pruvost[1] and Pierre Lefeuvre[1]

[1] UMR PVBMT, Cirad, F-97410, St Pierre, Réunion, France
[2] Plant Health Laboratory, ANSES, F-97410, St Pierre, Réunion, France
[3] Université de la Réunion, F-97410, St Pierre, Réunion, France

Corresponding Author: damien.richard@cirad.fr

Horizontal gene transfer (HGT) is often reported as being the motor of bacterial adaptation. Among the various processes involved in HGT, plasmid transfers were shown to be of prime importance for many bacterial families [1,2]. Plasmid-encoded accessory genes enrich the species' gene pool and provide new adaptive traits in response to environmental modifications, such as the use of new antimicrobials or the colonization of a new ecological niche [3,4]. Therefore, in order to further our understanding of bacterial adaptation and ecology, it is pivotal to improve our knowledge on (1) plasmid genomic structure, (2) mechanisms involved in plasmid transfers and (3) their limits. Phylogenomic networks are usefull tools to picture HGT among prokaryotes [1]. Indeed, each edge of the network represent an HGT event between two genomes (i.e. nodes of the network). Using this approach, it was demonstrated that HGT is limited by the taxonomic relatedness of the organisms, the characteristics of the host genomes (such as GC content or codon usage) and the ecological niche they occupy.

Copper-based antimicrobial compounds are widely used to control plant bacterial pathogens. *Xanthomonas citri* pv. *citri* (Xcc), a major citrus pathogen, has adapted in response to this selective pressure [5,6] and it was demonstrated that its copper-resistance was plasmid-borne. In order to understand the genetic basis of this adaptation, we fully sequenced 13 copper-resistant strains from the Xanthomonadaceae family: six Xcc strains [7], six *Xanthomonas* strains pathogenic to solanaceous species [8] and one commensal *Stenotrophomonas* strain [9]. Genome comparison between the six Xcc strains revealed that the copper-resistance genes were encoded on an adaptive transposon located on a conjugative plasmid. Most strains of other *Xanthomonas* species encoded a highly conserved copy of the entire plasmid, but one that displayed a similar adaptive transposon inserted in a distinct plasmid. The commensal *Stenotrophomonas* strain presented with a chromosomally encoded copy of the transposon. These findings highlighted the existence of two overlapping levels of mobility (transposon and plasmid) [10].

To further investigate the spread of copper-resistance genes among the Xanthomonadaceae family, we searched for homologues of every plasmid's genes in the public NCBI databases NR and WGS. Based on sequence identity, we constructed and analysed phylogenomic networks. It revealed that homologues of the plasmid present in Xcc were only identified from Xcc and few *Xanthomonas* species pathogenic to solaneous species. In contrast, genes homologous to the adaptive transposon were detected from 14 species included in five genera (*Xanthomonas*, *Stenotrophomonas*, *Pseudoxanthomonas*, *Pelomonas* and *Pseudomonas*) and grouping in three families. Globally, homologues of genes encoded on the transposon were found further apart in the taxonomy than plasmid backbone homologues.

A similar approach using all available plasmids from the Xanthomonadaceae family was applied to more finely define the reach of gene exchange for bacteria in this family. The analyses of the resulting networks would certainly help uncover the global context of gene exchange, their limits and specificities. It may also reveal specific trends for genes involved in adaptation within the particular context of agricultural settings.

References

[1] O. Popa and T. Dagan. Trends and barriers to lateral gene transfer in prokaryotes. *Curr Opin Microbiol*, (14):615-623, 2011.
[2] S. Halary, J. W. Leigh, B. Cheaib, P. Lopez and E. Bapteste. Network analyses structure genetic diversity in independent genetic worlds. Proceedings of the National Academy of Sciences, USA, (107):127-132, 2010.
[3] J. L. Hobman and L. C. Crossman. Bacterial antimicrobial metal ion resistance. *J Med Microbiol*, (64):471-497, 2015.
[4] H. Ochman, J. G. Lawrence and E. A. Groisman. Lateral gene transfer and the nature of bacterial innovation. *Nature*, (405):299-304, 2000.

[5] D. Richard, C. Boyer, S. Javegny, K. Boyer, P. Grygiel, O. Pruvost, A. L. Rioualec, V. Rakotobe, J. Iotti, R. Picard, C. Vernière, C. Audusseau, C. François, V. Olivier, A. Moreau and A. Chabirand. First Report of Xanthomonas citri pv. citri Pathotype A Causing Asiatic Citrus Canker in Martinique, France. *Plant Disease*, (100):1946-1946, 2016.

[6] D. Richard, N. Tribot, C. Boyer, M. Terville, K. Boyer, S. Javegny, M. Roux-Cuvelier, O. Pruvost, A. Moreau, A. Chabirand and C. Vernière. First Report of Copper-resistant Xanthomonas citri pv. citri Pathotype A Causing Asiatic Citrus Canker in Réunion, France. *Plant Disease*, (101):503, 2016.

[7] D. Richard, C. Boyer, C. Vernière, B. I. Canteros, P. Lefeuvre and O. Pruvost. Complete Genome Sequences of Six Copper-Resistant Xanthomonas citri pv. citri Strains Causing Asiatic Citrus Canker, Obtained Using Long-Read Technology. *Genome Announcements*, (5):2017.

[8] D. Richard, C. Boyer, P. Lefeuvre, B. I. Canteros, S. Beni-Madhu, P. Portier and O. Pruvost. Complete Genome Sequences of Six Copper-Resistant Xanthomonas Strains Causing Bacterial Spot of Solaneous Plants, Belonging to X. gardneri, X. euvesicatoria, and X. vesicatoria, Using Long-Read Technology. *Genome Announcements*, (5):2017.

[9] D. Richard, C. Boyer, P. Lefeuvre and O. Pruvost. Complete Genome Sequence of a Copper-Resistant Bacterium from the Citrus Phyllosphere, Stenotrophomonas sp. Strain LM091, Obtained Using Long-Read Technology. *Genome Announcements*, (4):2016.

[10] D. Richard, V. Ravigné, A. Rieux, B. Facon, C. Boyer, K. Boyer, P. Grygiel, S. Javegny, M. Terville, B. I. Canteros, I. Robène, C. Vernière, A. Chabirand, O. Pruvost and P. Lefeuvre. Adaptation of genetically monomorphic bacteria: evolution of copper resistance through multiple horizontal gene transfers of complex and versatile mobile genetic elements. *Molecular Ecology*, Special Issue: Microbial Local Adaptation, 2017.

Méthodes et outils de construction de super-arbres en phylogénie

Morgan SOULIÉ[2], Vincent LEFORT[1] et Anne-Muriel Arigon CHIFOLLEAU[1]
[1] LIRMM, UMR 5506, CNRS et Université Montpellier, 161 rue Ada, 34095, Montpellier, France
[2] Parcours Bioinformatique, Connaissances, Données (BCD) du Master Sciences et Numérique pour la Santé (SNS)

Auteur référent: morgan.soulie@etu.umontpellier.fr

Résumé

Un super-arbre est un arbre phylogénétique construit à partir d'une collection d'arbres, dits arbres sources, partageant complètement ou partiellement un même ensemble d'espèces. Les informations portées par chaque arbre source peuvent être complémentaires et contradictoires entre elles. Un super-arbre représente une synthèse d'hypothèses relationnelles qui couvre l'ensemble des organismes présents dans chacun des arbres sources. Il permet de révéler des liens de parenté non visibles dans les arbres sources pris séparément. Afin de résoudre les problèmes de compatibilité entre les parties communes des arbres sources, de nombreuses méthodes de construction de super-arbre ont été développées. La plateforme ATGC (www.atgc-montpellier.fr) de l'IFB (Institut Français de Bioinformatique), adossée à l'équipe MAB (Méthodes et Algorithmes pour la Bioinformatique) du LIRMM, est dédiée à la bioinformatique pour la génomique évolutive comparative et fonctionnelle. Nous souhaitons développer un portail web convivial dédié aux analyses phylogénomiques intégrant les outils de référence du domaine, et notamment les outils de construction de super-arbres. Ce portail permettra de guider les biologistes dans leurs analyses afin de leur fournir des services adaptés à leurs problématiques. Pour ce faire, nous avons réalisé un état des lieux des méthodes et outils de construction de super-arbre existants et identifié les plus utilisés. Nous avons sélectionné ceux qui pourraient être directement intégrés à la plateforme.

Plusieurs approches sont utilisées dans la construction de super-arbres : l'approche libérale ou de "vote" qui résout les conflits en "faisant voter" les arbres sources et en optant pour l'alternative topologique qui maximise un critère d'optimisation, celui-ci variant d'une méthode à l'autre (MRP - représentation matricielle avec parcimonie, similarité, bipartitions, quadruplets, SPR - subtree pruning and regrafting, ...) ; l'approche consensus qui utilise des méthodes issues de la théorie des graphes, l'approche "veto" dont la topologie des super-arbres inférés respectent la topologie de chacun des arbres sources ; ou de nouvelles approches basées, par exemple, sur le maximum de vraisemblance ou sur l'analyse des relations entre paires de taxons. Les outils actuels se basent sur ces approches et implémentent certaines de ces méthodes. Notre étude nous a permis de préselectionner un ensemble d'outils implémentant les principales méthodes : PhySIC IST (Scornavacca et al., 2008 ; approche libérale et approche "veto"), L.U.St (Akanni et al., 2014 ; approche par maximum de vraisemblance), COSPEDTree (Bhattacharyya et al., 2016 ; approche par paire de taxons), SPRsupertrees (Whidden et al., 2014 ; approche libérale) et Clann (Creevey et al., 2005 ; approche libérale et approche consensus). Parmi les méthodes de super-arbre, la méthode MRP est la méthode la plus populaire (Olaf R.P. Bininda-Emonds, 2004), et les articles la décrivant (Baum, 1992 et Ragan, 1992) ont respectivement été cités 459 et 529 fois. Par la suite, nous porterons donc une attention particulière à l'implémentation et l'utilisation de cette méthode.

Remerciements

Cette participation à la 18ème dition des Journes Ouvertes en Biologie, Informatique et Mathmatiques (JOBIM) a t finance par le Labex Numev (http ://www.lirmm.fr/numev/).

Implémentation d'une interface pratique pour l'évaluation de stratégies d'exploration de la diversité moléculaire par protéogénomique

YANNICK COGNE[1], CHRISTINE ALMUNIA[1], OLIVIER PIBLE[1], DUARTE GOUVEIA[2], ARNAUD CHAUMOT[2], OLIVIER GEFFARD[2] and JEAN ARMENGAUD[1]

[1] CEA-Marcoule LI2D, PRAE Marcel Boiteux, F-30200, Bagnols-sur-Cèze, France
[2] IRSTEA - Centre de Lyon, 5 rue de la Doua, 69616, Villeurbanne, France

Corresponding Author: jean.armengaud@cea.fr

Le projet ANR Proteogam propose de définir des marqueurs moléculaires d'un organisme sentinelle afin de tester rapidement son état de santé pour sonder la qualité biologique de l'eau d'une rivière. L'organisme sentinelle choisi est le gammare, un petit amphipode d'eau douce, présent dans tous les cours d'eau d'Europe. Les marqueurs moléculaires sont des protéines dont l'abondance peut varier en fonction de la présence de polluants dans les eaux de rivière. Typiquement, les concentrations de protéines présentes dans les gonades peuvent signer la présence de perturbateurs endocriniens.

Le Gammare est une espèce dite « non-modèle » car il n'existe, pour cette espèce, pas de séquence génomique de référence permettant l'attribution des spectres MS/MS enregistrés en protéomique. Les génomes de référence les plus proches sont ceux de *Hyalella azteca* et *Daphnia pulex* qui restent cependant trop éloignés pour une attribution suffisante. Pour pallier ce manque d'information, une base de données de référence peut être constituée par assemblage *de novo* de lectures de transcrits matures. Pour sonder la diversité des populations de *Gammarus fossarum*, la stratégie protéogénomique mise en œuvre nécessite la combinaison des outils bioinformatiques d'assemblage de données RNAseq et d'attribution protéomique [1]. Cette stratégie requiert un environnement de travail souple et reproductible, permettant la gestion simple et efficace de données volumineuses et de formats multiple.

L'environnement de travail choisi est Galaxy avec une solution portable aisée à l'aide de Docker, un programme de gestion de containers permettant son exécution sous Windows en mimant UNIX. L'interface proposée utilise donc Galaxy docker développé par Björn Grüning [2] .Une utilisation de docker dans le Galaxy docker a été choisie afin d'utiliser un container par outil. Cette séparation par outils permet via l'utilisation d'un gestionnaire d'installation (conda) d'apporter une gestion des versions et une mise à jour simple de chaque outil indépendamment ainsi qu'une forte reproductibilité des variables d'environnement au lancement de chaque outil.

La présentation détaillera l'implémentation pour l'évaluation de stratégies d'exploration de la diversité moléculaire par protéogénomique, notamment la réalisation de la gestion des données issues de nouvelles technologies de séquençage ARN et leurs assemblages.

References

[1] Armengaud, J., Trapp, J., Pible, O., Geffard, O., Chaumot, A., & Hartmann, E. M. (2014). Non-model organisms, a species endangered by proteogenomics. *Journal of proteomics*, *105*, 5-18
[2] Björn Grüning: https://github.com/bgruening/docker-galaxy-stable.
[3] Afgan, E., Baker, D., Van den Beek, M., Blankenberg, D., Bouvier, D., Čech, M., ... & Grüning, B. (2016). The Galaxy platform for accessible, reproducible and collaborative biomedical analyses: 2016 update. *Nucleic acids research*, gkw343.

Pea genetic map enrichment with Genotyping By Sequencing markers

Ayité Kougbeadjo[1], Grégoire Aubert[1], Mathieu Falque[2], David Edwards[3], Philipp Bayer[3], Jacqueline Batley[3], Manuel Zander[3], Satomi Hayashi[3], Jonathan Kreplak[1] and Judith Burstin[1]

[1] INRA UMR Agroecologie, 17 Rue Sully, 21000, Dijon, France
[2] INRA Ferme du Moulon, 91190, Gif sur Yvette, France
[3] School of biological Sciences , University of Western Australia

Corresponding Author: jonathan.kreplak@inra.fr

Pea (*Pisum sativum L.*) is a model plant and an important protein crop used in human and animal feed.

The laboratory developed a 13.2K SNP Infinium genotyping array and built individual and consensus genetic maps for [1] which gave information on the genome structure. We developed a pipeline on GBS (Genotyping by sequencing) data to enrich the consensus map with more markers to get a more accurate map.

New sequencing technologies as GBS (Genotyping by Sequencing) allow to access genetic information like markers cost-effectively. GBS is a simple and reproducible genome reduction approach using enzymes of restriction sensitive to methylation before sequencing. It allows reducing the number of sequence reads and can allow to avoid repetitive regions[2]. A Genotype by Sequencing (GBS) experiment was conducted by Warkentin et al. on a recombinant inbred line *Cameor x Melrose* produced by INRA.

Our pipeline first mapped the reads on the temporary reference assembly using bwa mem[3] and kept unique mapping with MAPQ (PHRED MAPping Quality) higher or equal to 30. Optical duplicates were removed with the picard tools[4]. We then called the variants using SAMtools mpileup and BCFtools[5]. The results showed that abnormal numbers of heterozygous were detected for all sites where the depth of sequencing was less than 3 reads.

To support low depth sites, we first called SNPs from the 7x resequencing data of the parents. We filtered them to only keep SNPs that are non-variant between *Cameor* resequencing reads and the *Cameor* genome assembly, and variant between *Melrose* resequencing reads and the *Cameor* genome assembly. This defines the set of possible SNPs in the progeny. Only GBS polymorphic sites that are common with parents polymorphic sites are taken into account. All GBS sites with a read depth below 3 was then called again using an in-house Python script.

We finally obtained a 473 587 SNPs of which 468 448 SNPs were placed on the genetic map. This map combined with a bionano map and a whole genome profiling map will allow us to rearrange scaffolds in our assembly and to do further analysis. We are working to encapsulate the pipeline in a reusable package.

References

1. Tayeh N, Aluome C, Falque M, Jacquin F, Klein A, Chauveau A, et al. Development of two major resources for pea genomics: the GenoPea 13.2K SNP Array and a high-density, high-resolution consensus genetic map. Plant J,(84): 1257–1273, 2015
2. Elshire RJ, Glaubitz JC, Sun Q, Poland JA, Kawamoto K, Buckler ES, et al. A Robust, Simple Genotyping-by-Sequencing (GBS) Approach for High Diversity Species. Public Library of Science, 2011
3. Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics, 2009
4. Picard [Internet]. Available: http://picard.sourceforge.net/
5. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, et al. The Sequence Alignment/Map format and

Développement d'une mesure du biais d'usage des codons : application aux virus humains

Jérôme BOURRET[1,2], Samuel ALIZON[1] et Ignacio G BRAVO[1]

[1] MIVEGEC(UMR CNRS 5290, IRD 224, UM), 911 avenue Agropolis, 34394, Montpellier, France
2 Parcours Bioinformatique, Connaissances, Données (BCD) du Master Sciences et Numérique pour la Santé (SNS), place Eugène Bataillon, 34090, Montpellier, France

Adresse mail : bourret.jerome@gmail.fr

Au cours de la synthèse d'une protéine à partir d'un ARN messager, la machinerie ribosomale traduit 61 codons en 20 acides aminés. Le code génétique est de ce fait « dégénéré » : seuls 2 acides aminés sont codés par des codons uniques alors que les 18 autres sont codés par des combinaisons de 2, 3, 4 ou 6 codons. Même si ces derniers sont généralement appelés « codons synonymes », leur utilisation ne se fait pas de manière aléatoire. Les préférences d'usage de certains codons varient entre *organismes,* entre *gènes à l'intérieur d'un génome* et même entre *les positions d'un même gène*. On parle de *biais d'usage des codons* [1]. Celui-ci peut être dû à un *biais mutationnel* : lors de la synthèse ou de la réparation de l'ADN, tous les nucléotides ne sont pas utilisés de manière identique, ce qui peut influencer la composition nucléotidique et donc biaiser l'usage du code [2]. Le second moteur du biais d'usage des codons est la *sélection traductionnelle* : elle repose sur l'hypothèse que certains codons confèrent une traduction plus efficace lors de la synthèse protéique [2].

Pour estimer le degré d'usage non-aléatoire des codons synonymes, plusieurs mesures ont été développées depuis les années 1980. Ces mesures appartiennent principalement à deux classes : l'une où les fréquences des codons synonymes d'une séquence requête sont comparées à celles d'un jeu de données de référence [3] et l'autre où l'hypothèse nulle correspond à un usage aléatoire des différents codons synonymes [4].

Notre objectif est de développer une nouvelle mesure qui combine ces deux approches : elle calcule le biais d'usage des codons d'une séquence requête face aux préférences d'usage des codons dans un jeu de données de référence tout en normalisant selon une hypothèse d'usage aléatoire des codons. Nous avons effectué le développement de cette mesure et l'avons implémentée au sein d'un outil codé en Python, R et C++. Nous avons ensuite réalisé une étude comparative entre notre mesure et les autres outils et mesures de biais d'usage des codons.

Dans un second temps, notre outil sera utilisé pour mesurer le biais d'usage des codons d'un large panel de virus humains en utilisant comme référence celui de leur hôte humain. D'éventuelles différences de biais d'usage des codons détectées par notre outil permettront d'aborder des phénomènes tels que le contrôle spatio-temporel différentiel de l'expression génique à l'intérieur d'un même génome ou l'exposition différentielle au système immunitaire des protéines virales en fonction des interactions hôte-parasite [5].

Remerciements

Cette participation à la 18ème édition des **JOBIM** a été financée grâce au laboratoire d'excellence Numev (http://www.lirmm.fr/numev/).

Références

[1] R. Grantham, C. Gautier, M. Gouy, R. Mercier and A. Pavé. *Codon catalog usage and the genome hypothesis*. Nucleic Acids Research 8(1) :r49–r62, 1980.
[2] A. Roth, M. Anisimova et G. M. Cannarozzi. *Measuring codon usage bias*. In: Cannarozzi G.M, Schneider A, editors. Codon Evolution Mechanism and Models:University of Bern.University of Utrecht:189-198, 2012.
[3] P. M. Sharp and W. H. Li. *The codon Adaptation Index–a measure of directional synonymous codon usage bias, and its potential applications*. Nucleic Acids Research 15(3) :1281–1295, 1987.
[4] F. Wright. *The 'effective number of codons' used in a gene*. Gene 87(1) :23–29, 1990.
[5] M. Félez-Sánchez, J. Trösemeier, S. Bedhomme, M.I. González-Bravo, C. Kamp, and I.G. Bravo. *Cancer, Warts, or Asymptomatic Infections: Clinical Presentation Matches Codon Usage Preferences in Human Papillomaviruses*. Genome Biology and Evolution, vol. 7, pp. 2117-2135, 2015.

Was the Chlamydial Adaptative Strategy to Tryptophan Starvation an Early Determinant of Plastid Endosymbiosis?

Ugo CENCI[1], Mathieu DUCATEZ[1], Derifa KADOUCHE[1], Christophe COLLEONI[1] and Steven G. BALL[1]

[1] *Unité de Glycobiologie Structurale et Fonctionnelle, UMR8576 Centre National de la Recherche Scientifique, Université des Sciences et Technologies de Lille, Villeneuve d'Ascq, France*

[2] Laboratory, Address, zip code, Town, Country

Corresponding Author: steven.ball@univ-lille1.fr

Abstract

Chlamydiales were recently proposed to have sheltered the future cyanobacterial ancestor of plastids in a common inclusion [1,2]. The intracellular pathogens are thought to have donated those critical transporters that triggered the efflux of photosynthetic carbon and the consequent onset of symbiosis [3,4]. Chlamydiales are also suspected to have encoded glycogen metabolism TTS (Type Three Secretion) effectors responsible for photosynthetic carbon assimilation in the eukaryotic cytosol [1]. We now turn our attention to the reasons underlying other chlamydial lateral gene transfers evidenced in the descendants of plastid endosymbiosis. In particular, we show that half of the genes encoding enzymes of tryptophan synthesis in Archaeplastida are of chlamydial origin. Tryptophan concentration is an essential cue triggering two alternative modes of replication in Chlamydiales. In addition, sophisticated tryptophan starvation mechanisms are known to act as antibacterial defenses in animal hosts [5,6]. We propose that Chlamydiales have donated their tryptophan operon to the emerging plastid to ensure increased synthesis of tryptophan by the plastid ancestor. This would have allowed massive expression of the tryptophan rich chlamydial transporters responsible for symbiosis. It would also have allowed possible export of this valuable amino-acid in the inclusion of the tryptophan hungry pathogens. Free-living single cell cyanobacteria are devoid of proteins able to transport this amino-acid. We therefore investigated the phylogeny of the Tyr/Trp transporters homologous to *E. coli* TyrP/Mre and found yet another LGT from Chlamydiales to Archaeplastida thereby considerably strengthening our proposal.

References

[1] Ball, S.G., Subtil, A., Bhattacharya, D., Moustafa, A., Weber, A.P.M., Gehre, L., Colleoni, C., Arias, M.-C., Cenci, U., and Dauvillée, D. Metabolic Effectors Secreted by Bacterial Pathogens: Essential Facilitators of Plastid Endosymbiosis? *Plant Cell* 25, 7–21, 2013.

[2] Facchinelli, F., Colleoni, C., Ball, S.G., and Weber, A.P.M. Chlamydia, cyanobiont, or host: who was on top in the ménage à trois? *Trends Plant Sci.* 18, 673–679, 2013.

[3] Facchinelli, F., Pribil, M., Oster, U., Ebert, N.J., Bhattacharya, D., Leister, D., and Weber, A.P.M. Proteomic analysis of the Cyanophora paradoxa muroplast provides clues on early events in plastid endosymbiosis. *Planta* 237, 637–651, 2013.

[4] Karkar, S., Facchinelli, F., Price, D.C., Weber, A.P.M., and Bhattacharya, D. Metabolic connectivity as a driver of host and endosymbiont integration. *Proc. Natl. Acad. Sci. 112*, 10208–10215, 2015.

[5] Bonner, C.A., Byrne, G.I., and Jensen, R.A. Chlamydia exploit the mammalian tryptophan-depletion defense strategy as a counter-defensive cue to trigger a survival state of persistence. *Front. Cell. Infect. Microbiol.* 4, 2014.

[6] Lo, C.-C., Xie, G., Bonner, C.A., and Jensen, R.A. The AlternativeTranslational Profile That Underlies the Immune-Evasive State of Persistence in Chlamydiaceae Exploits Differential Tryptophan Contents of the Protein Repertoire. *Microbiol. Mol. Biol. Rev.* 76, 405–443, 2012.

OrthoInspector 3.0: orthology en route to big data

Yannis NEVERS[1], Arnaud KRESS[1], Raymond RIPP[1], Olivier POCH[1] and Odile LECOMPTE[1]

[1] CSTB team, UMR7357 - ICube, 4 rue Kirschleger 67085 Strasbourg, France

Corresponding Author: yannis.nevers@etu.unistra.fr

Homologs, genes that derive from a common ancestor, can be separated in two classes: paralogs -deriving from an ancestral gene by a duplication event, and orthologs - deriving from an ancestor by a speciation event. Orthologs generally perform a similar function in different species with possible species-specific adaptations. Prediction of correct orthology relationship is critical in many functional and evolutionary applications: phylogenetic tree inferences, gene function predictions, genome annotation, selection of a relevant model organism for experimental studies... OrthoInspector [1] is one of the leading algorithm [2] for pairwise orthologous relationship predictions. In addition to the independent software package, the OrthoInspector website (www.lbgi.fr/orthoinspector) currently offers precomputed databases, with a sampling of 259 eukaryotes, 1568 bacteria and 120 archaea.

Here, we present our current developments to build a new release of OrthoInspector databases. This major update aims at presenting a comprehensive set of species. As a starting point, we used a selection of non-redundant proteome representing a wide variety of taxa: the Uniprot Reference Proteomes (Nov 2012) [3]. Statistical analyses of proteome contents were used to identify and filter low-quality proteomes. The 4752 selected proteomes (representing 87% of the original set) were used as our dataset for the BLAST all-versus-all [4] searches, first step of OrthoInspector calculation. With more than 20 million BLASTP searches, this step is computationally expensive. Thus, we segmented the BLAST database to parallelize this step and to perform it on the EGI computing grid [5]. Computing of orthology relationships were in turn parallelized on our local infrastructure. The final computed databases of orthology relationships provide a major breakthrough in terms of covered species with 711 eukaryotes, 3862 bacteria and 179 archaea. It makes OthoInspector database the most comprehensive orthology database available to date.

In parallel, we are developing new features to the OrthoInspector suite to improve the possibilities offered by OrthoInspector package and website. This includes support for SQLite databases, automatic update procedures to keep our resources up-to-date with Uniprot Reference Proteomes and more importantly, a definition of ortholog families. Families will be available in complement to the pairwise relationships currently supported. We plan to provide evolutionary characterization of these gene family through multiple sequence alignment, domain conservations and integration of cross-references toward external biological resources.

References

[1] B. Linard *et al.*, "OrthoInspector 2.0: Software and database updates," *Bioinforma. Oxf. Engl.*, vol. 31, no. 3, pp. 447–448, Feb. 2015.
[2] A. M. Altenhoff *et al.*, "Standardized benchmarking in the quest for orthologs," *Nat. Methods*, vol. 13, no. 5, pp. 425–430, May 2016.
[3] UniProt Consortium, "UniProt: a hub for protein information," *Nucleic Acids Res.*, vol. 43, no. Database issue, pp. D204-212, Jan. 2015.
[4] S. F. Altschul, W. Gish, W. Miller, E. W. Myers, and D. J. Lipman, "Basic local alignment search tool," *J. Mol. Biol.*, vol. 215, no. 3, pp. 403–410, Oct. 1990.
[5] European Grid Infrastructure. http://www.egi.eu

Déploiement automatique d'une infrastructure complexe pour le mapping des données de séquençage

Sandrine PERRIN [1], Bryan BRANCOTTE [1], Jonathan LORENZO [1], Christophe BLANCHET [1] and Jean-Francois GIBRAT [1]

[1] CNRS, UMS 3601 ; Institut Français de Bioinformatique, IFB-core, Avenue de la Terrasse, F-91190 Gif-sur-Yvette, France

Corresponding Author: sandrine.perrin@france-bioinformatique.fr

L'accroissement des données de NGS implique que les équipes disposent de solutions performantes, souples et reproductibles pour le traitement de données volumineuses. De telles infrastructures sont très lourdes pour être proposées au niveau local. L'Institut Français de Bioinformatique (IFB)[1] est une infrastructure nationale de service en bio-informatique pour les sciences de la vie. Elle met à disposition un cloud académique et anime un réseau national de plateformes bioinformatiques. Cette e-infrastucture de calcul est l'environnement adéquat pour répondre aux besoins de traitement des données biologiques. Elle met à disposition des machines virtuelles pré-configurées, des clusters sur-mesure et prochainement des gestionnaires de données volumineuses dont les données de référence.

Le projet européen CYCLONE H2020 [2] a pour mission d'aider aux développements de solutions pour déployer facilement des infrastructures complexes, il apporte notamment des utilitaires pour la construction de VPN et une authentification sécurisée.

La solution présentée ici répond à un cas d'utilisation proposé par CYCLONE sur une solution prenant en charge l'étape de mapping, qui est consommatrice en ressources mémoire (chargement des index), en ressources de calcul et surtout en temps (taille des données d'entrée). Cette étape est préalable à tout traitement sur les données de séquençage. Elle est facilement automatisable. La solution développée par l'IFB propose un déploiement en un clic d'un cluster configurable en fonction des besoins de l'analyse à réaliser et prêt à l'emploi, elle comprend :

- un cluster swarm docker [3] extensible et configurable ;
- un accès à un gestionnaire de données (tel que iRODS), performant avec des volumes importants, à la charge de l'utilisateur de télécharger ces données dans son espace sur le cloud ;
- une interface graphique de gestionnaire de conteneurs Docker, l'analyse sera faite par des outils dockerisés référencés, par exemple dans BioShadock [4], BioContainer ou Docker-hub. Dans notre cas, l'outil TopHat2 [5] est utilisé pour comparer les temps de traitement avec les résultats obtenus classiquement. D'autres étapes peuvent être intégrées : par exemple un pipeline de contrôle qualité ;
- un accès à une banque de données de référence gérée avec BioMaj [6] ;
- un réseau virtuel privé (VPN) pour isoler le cluster et l'accès aux données, la solution CNSMO a été développée dans le cadre du projet CYCLONE ;
- un accès sécurisé au cluster grâce à la fédération d'identités eduGAIN pour les accès SSH et Web ;
- au terme de l'étape de mapping, l'analyse peut être poursuivie avec d'autres outils mis à disposition dans le cloud au sein du cluster ou dans d'autres machines virtuelles, disponibles dans le catalogue RAINBio [7].

L'IFB-core travaille parallèlement à la mise en place d'une fédération de clouds avec les plateformes bioinformatiques françaises partenaires. Les utilisateurs pourront alors rapprocher au mieux les environnements de calcul de leurs données.

[1] Institut Français de Bioinformatique : http://www.france-bioinformatique.fr/
[2] Projet CYCLONE (EU H2020 – 644925), http://www.cyclone-project.eu/
[3] Docker & Swarm : https://docs.docker.com/engine/swarm/ ; https://www.docker.io/
[4] Moreews F, Sallou O, Ménager H *et al*. BioShaDock: a community driven bioinformatics shared Docker-based tools registry. *F1000Research* 4:1443 2015.
[5] Kim D, Pertea G, Trapnell C, Pimentel H, Kelley R, Salzberg SL. TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. *Genome Biology* (14:R36) 2013.
[6] BioMAJ, http://biomaj.genouest.org
[7] RAINBio, https://biosphere.france-bioinformatique.fr/catalogue/

State of the art and comparison of long reads technologies

Anais POIRAUDEAU[1], Maxime MANNO[1], Céline VANDECASTEELE[1], Alain ROULET[1], Celine ROQUES[1], Marie VIDAL[1], Catherine ZANCHETTA[1], Pauline HEUILLARD[1], Fabrice ROUX[2], Baptiste MAYJONADE[2], Jérôme GOUZY[2], Yann GUIGUEN[3], Christophe KLOPP[4], Pierre FRASSE[5], Mohamed ZOUINE[5], Cécile DONNADIEU[1], Olivier BOUCHEZ[1], Gérald SALIN[1] and Claire KUCHLY[1]

1 INRA, US 1426, GeT-PlaGe, Genotoul, Castanet-Tolosan, France
2 LIPM, UMR 441 INRA, 31326, Castanet-Tolosan, France
3 INRA, UR 1037 LPGP Laboratoire de Physiologie et Génomique des Poissons, 35000, Rennes, France.
4 MIAT, UR875 INRA, Plateforme bioinformatique, 31326, Castanet-Tolosan, France
5 ENSAT, UMR 990 INRA/INPT-ENSAT, Laboratoire Génomique et Biotechnologie des Fruits, 31326, Castanet-Tolosan, France

Corresponding Author: maxime.manno@inra.fr

Genomic issues such as complex genome assembly, structural variant discovery or phasing can be addressed by Long Read technologies.

Thanks to its experience on short reads sequencing using the Illumina technologies, the GeT-PlaGe core facility began to evaluate and use long read technologies since the beginning of 2015 : Pacific BioSciences RSII, Oxford Nanopore Technology MinION and 10XGenomics Chromium. Results from different genomes sequenced by those platforms and sequences obtained on a PacBio Sequel from another France Genomic core facility will be presented and compared on this poster.

As DNA quality is the most important requirement to obtain an efficient sequencing, sample requirements for each technology, and quality controls performed on GeT-PlaGe will be detailed in a first point. Secondly, we will compare sequencing data in terms of file format, sequence metrics, barcoding solutions and needed IT resources. Current projects will also be presented concerning de novo assembly results obtained using Long Read technologies, for several genomes (bacteria, plant and fish).

Acknowledgements

We thank the GenoToul bioinformatics facility for its support in computing resources and data storage, and for helping us processing the data and making available all the needed software and infrastructure.

Genomic markers of species diversification in vertebrates

Guillaume LOUVEL[1], Eric LEWITUS[1], Hélène MORLON[1] and Hugues ROEST CROLLIUS[1]
Institut de biologie de l'École Normale Supérieure, 46 rue d'ulm, 75005, Paris, France

Corresponding author: guillaume.louvel@ens.fr

Abstract

Several mechanisms have been hypohesized to explain species divergence, from genomic incompatibilities to divergent selection pressures [1]. Given the current availability of full genomes for many non-model organisms sampling various branches of the vertebrate phylogeny, we can now combine genomic data with patterns of speciation from more complete phylogenies [2]. As Ohno [3] initially postulated, duplicated genes are good candidates for generating functional novelty and adaptation. We first dated duplications using dS (synonymous substitution rate) calculations, in order to obtain a fine estimation of the rate of gene duplication through time and lineages. This method however is sensitive to multiple bias (fast evolving branches, quality of gene alignments, etc). We are currently working on improving these dating method, and aim towards more sophisticated models of gene evolution that could estimate duplication ages (either adapting existing models or developping one). Our broader aim is to compare duplication rates with the diversification rates of taxons, and to assess the role of gene duplication in evolution.

References

[1] J A Coyne and H A Orr. The evolutionary genetics of speciation. *Philosophical transactions of the Royal Society of London. Series B, Biological sciences*, 353(1366):287–305, feb 1998.

[2] Eric Lewitus and Hélène Morlon. Natural Constraints to Species Diversification. *PLOS Biology*, 14(8):e1002532, aug 2016.

[3] S Ohno. *Evolution by Gene Duplication*. Springer-Verlag, 1970.

PREMS / ELVIS : A local plant biological resource management system

Fabrice DUPUIS[1], Aurélie LELIÈVRE[1], Sandra PELLETIER[1], Tatiana THOUROUDE[1], Julie BOURBEILLON[1] and Sylvain GAILLARD[1]

[1]IRHS, Agrocampus-Ouest, INRA, Université d'Angers, SFR 4207 QuaSaV, 49071, Beaucouzé, France

Corresponding Author: sylvain.gaillard@inra.fr

The "Institut de Recherche en Horticulture et Semences" (IRHS) hosts some collections of biological resources for bacteria, fungi and plants. These collections are part of certified Biological Resource Centers (BRC): the "Collection Française de Bactéries associées aux Plantes" (CFBP) for microbes and more recently the "Collection FRuits À PÉpins et Rosiers" (FrAPeR) and the "Collection Apiaceaes" for the plants. The management of these last two has leaded to the development of a tool for tracking the material and logging phenotyping and characterization of these resources.

We have developed a webservice oriented database (ELVIS [1]: Expérimentations en Laboratoire Végétal Information System) coupled with a dynamic web interface (PREMS [2]: Plant REsources Management System). These tools take care of specificities linked to the material maintenance (seeds for annual / biannual plants or vegetative reproduction for perennial plants) as well as confidentiality for breeding program.

Featured functionalities:

- Orchard / field location (apple tree, rose...) or geolocation (wild material characterization) through multi-referential tagging

- Material characterization (plant phenotyping in field, greenhouse, laboratory)

- Material input / output tracking

- Material sample and exploitation tracking (sample, location, transformation)

- Document association (MTA, protocols)

As a webservice oriented application, ELVIS can implement API to communicate with other systems like SIReGal [3], the national portal of INRA for plant genetic resources and through this be a part of the European network for genetic resource management. Moreover, the local installation of PHIS, the PHENOME [4] information system, in the laboratory allows us to build interoperability bridges.

ELVIS is built with Python and PostgreSQL and is freely available under open source CeCILL license. The webservice API uses the JSON-RPC protocol.

PREMS is powered by Qooxdoo [5] and also available under open source CeCILL license.

References

[1] https://sourcesup.renater.fr/projects/elvis/
[2] https://sourcesup.renater.fr/projects/prems/
[3] http://prodinra.inra.fr/record/216501
[4] https://www.phenome-fppn.fr/
[5] http://www.qooxdoo.org/

Découverte et analyse de polymorphismes SNPs issus de RNA-seq chez le peuplier noir

Odile ROGIER[1], Souhila AMANZOUGARENE[1], Marie-Claude LESAGE-DESCAUSES[1], Sandrine BALZERGUE[2,3], Véronique BRUNAUD[2,3], José CAIUS[2,3], Aurélien CHATEIGNER[1], Ludivine SOUBIGOU-TACONNAT[2,3], Véronique JORGE[1] and Vincent SEGURA[1]

[1] INRA, UR0588, UAGPF Amélioration Génétique et Physiologie Forestières, F-45075 Orléans, France
[2] Institute of Plant Sciences Paris Saclay IPS2, CNRS, INRA, Université Paris-Sud, Université Evry, Université Paris-Saclay, Bâtiment 630, 91405 Orsay, France
[3] Institute of Plant Sciences Paris-Saclay IPS2, Paris Diderot, Sorbonne Paris-Cité, Bâtiment 630, 91405 Orsay, France

Corresponding Author: odile.rogier@inra.fr

Le peuplier noir (*Populus nigra*) est une espèce majeure de la forêt alluviale en Europe occidentale et un support de la biodiversité écosystémique. Pour explorer les composantes de la variabilité génétique au sein de cette espèce parentale de peupliers hybrides commerciaux, nous avons initié une approche intégrative combinant des données génomiques, transcriptomiques et phénotypiques dans une vaste collection d'individus échantillonnés à partir de populations naturelles et évalués dans un même site expérimental. De par leur grande abondance et facilité d'accès par séquençage, les SNPs (Single Nucleotide Polymorphisms) sont des marqueurs moléculaires de choix pour les études de génétique quantitative (incluant les études d'association), de génétique des populations ou de sélection génétique. Chez le peuplier noir, si les premières études de polymorphismes de séquence se sont focalisées sur le reséquençage de quelques gènes candidats [1,2,3] ou sur l'identification de variants rares [4], des travaux plus récents se sont intéressés à une analyse plus globale avec le développement d'une puce de génotypage à partir de SNPs détectés par séquençage de génomes complets [5]. Nous rapportons ici l'analyse de polymorphismes SNPs dans des séquences transcrites produites par RNA-seq. Plus précisément, nous présentons un pipeline de détection et de typage de SNPs à partir de données RNA-seq ainsi que sa validation grâce à des données déjà existantes.

L'ARN a été extrait à partir de jeune xylème et de cambium recueillis sur 24 arbres correspondant à 2 répétitions de 12 génotypes originaires de 6 populations naturelles représentatives de l'aire de distribution de l'espèce en Europe occidentale. Après quantification de l'ARN par tissu, les ARNs de xylème et de cambium du même arbre ont été mélangés et soumis à un séquençage à haut débit sur un HiSeq2000 d'Illumina (séquences pairées de 100 pb). Nous avons mis en place un pipeline de détection de SNPs à partir de données RNA-seq. Les séquences ont d'abord été nettoyées puis alignées sur le génome de référence (*Populus trichocarpa)*. Ensuite, nous avons effectué un post-traitement des données d'alignement et nous avons lancé la découverte de variants à l'aide 3 outils en plusieurs modes, combinés à plusieurs paramètres de filtration. Les résultats obtenus par ces différentes techniques ont été comparés. Enfin, les SNPs détectés par les différents outils ont été annotés. Pour valider la technique, nous avons comparé les génotypes trouvés à des résultats de génotypage issus d'une puce Illumina Infinium 12k fournissant 7 903 SNPs polymorphes et qui avait été utilisée pour génotyper une collection de peupliers noirs incluant nos 12 génotypes étudiés [5]. Le taux moyen de similarité de génotypage aux positions communes entre les SNPs de la puce et ceux trouvés par RNA-seq (d'un même individu) est supérieur à 98%. Ces résultats démontrent la faisabilité et l'intérêt du RNA-seq pour la découverte et le typage de SNPs dans des populations naturelles.

References

[1] Fabio Marroni *et al.* Large-Scale Detection of Rare Variants via Pooled Multiplexed next-Generation Sequencing: Towards next-Generation Ecotilling: Detection of Rare Variants via next-Generation Sequencing. *The Plant Journal* 67 (4): 736–45, 2011.

[2] Fabio Marroni *et al.* Nucleotide Diversity and Linkage Disequilibrium in *Populus Nigra* Cinnamyl Alcohol Dehydrogenase (CAD4) Gene. *Tree Genetics & Genomes* 7 (5): 1011–23, 2011.

[3] Fernando P. Guerra *et al.* Association Genetics of Chemical Wood Properties in Black Poplar (*Populus Nigra*). *New Phytologist* 197 (1): 162–76, 2013.

[4] Bartel Vanholme *et al.* Breeding with Rare Defective Alleles (BRDA): A Natural *Populus Nigra* HCT Mutant with Modified Lignin as a Case Study. *New Phytologist* 198 (3): 765–76, 2013.

[5] Patricia Faivre-Rampant *et al.* New resources for genetic studies in *Populus nigra* : genome-wide SNP discovery and development of a 12k Infinium array. *Molecular Ecology Resources* 16 (4): 1023–1036, 2016.

DiNAMO: Exact method for degenerate IUPAC motifs discovery, characterization of sequence-specific errors

Chadi SAAD[1,2], Laurent NOÉ[2], Hugues RICHARD[3], Julie LECLERC[1], Marie-Pierre BUISINE[1], Hélène TOUZET[2] and Martin FIGEAC[4]

[1] JPARC (UMR1172 Inserm, Lille 2 University and Lille University Hospital), 59000, Lille, France

[2] CRIStAL (UMR CNRS 9189 Lille 1 University and Inria Lille), Team BONSAI, 59000, Lille, France

[3] LCQB (UMR 7238 CNRS Pierre Marie Curie University), 75006, Paris, France

[4] Functional and Structural Genomic Platform, Lille 2 University, 59000, Lille, France

Corresponding author: Chadi.Saad@univ-lille1.fr

Next generation sequencing technologies are still associated with relatively high error rates, about 1%, which correspond to thousands of errors in the scale of a complete genome. Each region needs therefore to be sequenced several times and variants are usually filtered based on depth criteria. The significant number of artifacts, in spite of those filters, shows the limit of conventional approaches and indicates that some sequencing artifacts are recurrent. This recurrence underlines that sequencing errors can depend on the upstream nucleotide sequence context. Our goal is to search for overrepresented motifs that tend to induce sequencing errors.

Previous studies showed that some motifs, such as GGT [1,2], induce sequencing errors in the Illumina technologies. However, these studies were dedicated to exact motifs, and did not take into account approximate motifs, limiting the statistical power of such approaches. On the other hand, some tools, such as FIRE [3], DREME [4] and Discrover [5], were developed to search for degenerate motifs over the 15-letter IUPAC alphabet in the context of chip-seq studies. However, these tools use greedy algorithms, implying a lack of sensitivity. So we developed an exact algorithm to search for degenerate motifs by enumerating all possible IUPAC motifs. This algorithm is based on mutual information and uses hashtables with graphs data structure to store the motifs. It is independent from the sequencing technology.

Experimental results on real data show that there are many overrepresented motifs upstream of sequencing artifacts. These latter are identified through the strand bias between forward and reverse reads. The homopolymer of length 3 CCC seems to be sufficient to induce errors on IonTorrent. On Illumina, motifs are mainly composed of GGC followed by GGT (like: TGGCNGGT) or homopolymers. We have also noticed a base quality fall after the detected motifs. Our exact algorithm requires less than one minute (Intel® Core™ i5-4570 CPU, 3.20GHz), and less than 2GB of RAM to search for full degenerate motifs of length 6 on a dataset of approximately 24000 sequences, extracted from 11 exomes sequenced on IonTorrent Proton.

Availability: https://github.com/bonsai-team/DiNAMO

Acknowledgements

This work is supported by *Lille University Hospital* and *Hauts-de-France region*

References

[1] Manuel Allhoff, Alexander Schönhuth, Marcel Martin, Ivan G Costa, Sven Rahmann, and Tobias Marschall. Discovering motifs that induce sequencing errors. In *BMC bioinformatics*, volume 14, page S1. BioMed Central Ltd, 2013.

[2] Frazer Meacham, Dario Boffelli, Joseph Dhahbi, David IK Martin, Meromit Singer, and Lior Pachter. Identification and correction of systematic error in high-throughput sequence data. *BMC bioinformatics*, 12(1):451, 2011.

[3] Olivier Elemento, Noam Slonim, and Saeed Tavazoie. A universal framework for regulatory element discovery across all genomes and data types. *Molecular cell*, 28(2):337–350, 2007.

[4] Timothy L Bailey. Dreme: motif discovery in transcription factor chip-seq data. *Bioinformatics*, 27(12):1653–1659, 2011.

[5] Jonas Maaskola and Nikolaus Rajewsky. Binding site discovery from nucleic acid sequences by discriminative learning of hidden markov models. *Nucleic acids research*, 42(21):12995–13011, 2014.

Evolutionary conservation of unusual N-glycosylation sites in the human glycosyltransferase B4GALNT2

Virginie COGEZ[1], Anaïs BARRAY[2], Jérôme DE RUYCK[1], Sophie GROUX-DEGROOTE[1] and Anne HARDUIN-LEPERS[1]

1. Univ.Lille, CNRS, UMR 8576, Structural and Functional Glycobiology Unit, 59655 Villeneuve d'Ascq, France
2. Bilille, CRIStAL (UMR CNRS 9189 Univ. Lille), 59655 Villeneuve d'Ascq, France

Corresponding Author: virginie.cogez@univ-lille1.fr

Post translational modifications (PTM) enhance the functional diversity of proteins. Among these PTM, glycosylation represents one of the most abundant and complex modification of proteins and N-glycosylation is determinant for cell survival. For 96% of N-glycosylproteins, glycosylation occurs at the canonical motif [N]-!P-[S/T] known as *sequon* [1]. In the laboratory, we study the evolutionary relationships and molecular evolution of glycosyltransferases that are Golgi enzymes implicated in the last steps of glycosylation [2] like B4GALNT2 implicated in the synthesis of blood group antigen Sda [3]. This enzyme is a glycoprotein and our working hypothesis is that the gain or loss of an ancestrally conserved N-glycosylation site in this enzyme might have resulted in the evolution of protein structure, subcellular localization and functional modifications with impact on the phenotype [4-7]. As a first towards answering this question, we identified over 150 B4GALNT-related sequences (37 B4GALNT1 paralogues and 85 B4GALNT2 orthologues) from public databases (NCBI, Ensembl, CAZy) using a BLAST approach. We carried out sequence-based analysis using various multiple sequence alignment algorithms (MUSCLE, ClustalΩ, MAFFT, T-Coffee, ClustalW) to delineate informative regions and define conserved sequence motifs. Furthermore, we used NetNGlyc 1.0 server [8] and GlycoMine [9] to predict potential N-glycosylation sites. We found differentially conserved N-glycosylation sites in the homologous B4GALNT sequences: a canonical sequon, [N]-!P-[S/T], present in 11 mammalian B4GALNT2 sequences, which has disappeared in the human B4GALNT2 sequence [10]. This N-glycosylation site is also present in some ancestral B4GALNT sequences and has disappeared in the paralogous B4GALNT1 sequences suggesting no major role for this N-glycan. Interestingly, two atypical N-glycosylation sites [N]-X-[C] and [N]-[G] were predicted in mammalian B4GALNT2 sequences. Despite the fact that the [N]-[G] site is present in almost all the mammalian B4GALNT2 sequences, we found using PNGase F, an enzyme that remove N-Glycans, that the human B4GALNT2 enzyme is not glycosylated at this position, but at a second unusual [N]-X-[C] glycosylation site. This latter was found to be highly conserved in nearly all B4GALNT sequences except the teleostean B4GALNT2 sequences further suggesting an important functional role, which is currently investigated.

References

[1] Zielinska DF, Gnad F, Wiśniewski JR, Mann M. *Cell*. May 28;141(5):897-907, 2010

[2] Harduin-Lepers A. *Glycobiology Insights* 2, 29-61, 2010.

[3] Dall'Olio F, Malagolini N, Chiricolo M, Trinchera M, Harduin-Lepers A. *Biochim Biophys Acta*, 1840(1):443-53, 2014.

[4] Ruggiero FM, Vilcaes AA, Iglesias-Bartolomé R, Daniotti JL. *Biochem J.* Jul 1;469(1):83-95, 2015.

[5] Yi L, Bozkurt G, Li Q, Lo S, Menon AK, Wu H. *Sci Rep*. Apr 4;8:45912, 2017.

[6] Mühlenhoff M, Manegold A, Windfuhr M, Gotza B, Gerardy-Schahn R. *J Biol Chem.* Sep 7;276(36):34066-73, 2001.

[7] Zhuo Y, Yang JY, Moremen KW, Prestegard JH. *J Biol Chem*, 291(38):20085-95, 2016.

[8] R. Gupta, and S. Brunak. *Pacific Symposium on Biocomputing* 7:310-322, 2002

[9] Li, F, Li, C, Wang, M, Webb, GI, Zhang, Y, Whisstock, JC, Song J. *Bioinformatics* 31: 1411-1419, 2015

[10] Montiel, L, Krzewinski-Recchi, MA, Delannoy, P, Harduin-Lepers, A. *Biochem. J.* 373, 369-379. 2003

Comparison of statistical methods of inference of cooccurrence networks within microbial ecosystems from metagenomics data

Julie LAO[1], Mahendra MARIADASSOU[1] and Sophie SCHBATH[1]

[1] Institut National de la Recherche Agronomique, UR 1404 MaIAGE, F-78352, Jouy-en-Josas, France

Corresponding Author: julie.lao@inra.fr

Metagenomics consists of experimentally characterizing a microbial ecosystem as a whole without prior isolation of the different microorganisms composing it. Many microorganisms are not culturable and separate analyses of each microorganism result in a warped understanding of the ecosystem, as they overlook close relationships between these microorganisms (mutualism, parasitism). Metagenomics enable us to apprehend a microbial ecosystem globally.

Metagenomics data raise many methodological questions, as studies are increasingly moving beyond the mere constitution of a catalogue of species or genes and towards more complex analyses accounting for spatial data, time series and covariates. In particular, it is not clear how best to perform interaction studies and, more precisely, how to detect associations within the ecosystem [1].

In recent years, several statistical methods were developed to detect significant cooccurrences between species, in different ecosystems and under different experimental conditions. These methods assume that cooccurences are indicative of biological interactions between species [1] and interactions are thus revealed by reconstructing the cooccurrence network. SparCC [2], REBACCA [3] and SPIEC-EASI [4] are recently developed tools for the problem of network reconstruction in microbial ecology.

On this poster, we present a benchmark on the main reconstruction methods. Accuracy and running time was assessed on both simulated and real metagenomic data.

References

[1] Lima-Mendez, G., Faust, K., Henry, N., Decelle, J., Colin, S., Carcillo, F., … Raes, J. (2015). Determinants of community structure in the global plankton interactome. *Science*, *348*(6237), 1262073. doi:10.1126/science.1262073

[2] Friedman, J., & Alm, E. J. (2012). Inferring Correlation Networks from Genomic Survey Data. *PLOS Computational Biology*, *8*(9), e1002687. doi:10.1371/journal.pcbi.1002687

[3] Ban, Y., An, L., & Jiang, H. (2015). Investigating microbial co-occurrence patterns based on metagenomic compositional data. *Bioinformatics*, *31*(20), 3322-3329. doi:10.1093/bioinformatics/btv364

[4] Kurtz, Z. D., Müller, C. L., Miraldi, E. R., Littman, D. R., Blaser, M. J., & Bonneau, R. A. (2015). Sparse and Compositionally Robust Inference of Microbial Ecological Networks. *PLOS Computational Biology*, *11*(5), e1004226. doi:10.1371/journal.pcbi.1004226

ShRCAn : a user-friendly R-Shiny application for quantitative metagenomics analysis

Florence THIRION[1], Emmanuelle LE CHATELIER[1], Nicolas PONS[1], Anne-Sophie ALVAREZ[1], Pierre LEONARD[2], Dusko EHRLICH[1,3]

1. MGP MetaGénoPolis, Institut National de Recherche Agronomique, Université Paris-Saclay, 78350 Jouy en Josas, France.
2. DSI Global Services, 92350 Le Plessis-Robinson, France
3. Centre for Host Microbiome Interactions, Dental Institute, King's College London, UK.

Corresponding Author: florence.thirion@inra.fr

1. Introduction

In quantitative metagenomics, sequenced reads are usually mapped onto a reference gene catalogue yielding a gene abundance table (or counting matrix). The increasing number of metagenomic studies and reference gene catalogues makes it necessary to develop efficient new tools able to deal with those abundance tables, in order to ensure standardization of analysis as well as making it easier and faster for the user, especially concerning repeated tasks. Here we present ShRCAn, a user-friendly software that takes an abundance table as input to preprocess the data, perform the preliminary analysis, and enable data visualization.

2. Materials and methods

ShRCAn (Shiny application for Raw Counting matrix Analysis) is written in the free language R. It uses the package shiny, that helps in building interactive applications in R; the package MetaOMineR (momr)[1] that provides useful functions for analysis of metagenomic data; the package ggplot2 that creates graphics; and optionally the package PARConnector that enables to use the ProActive Parallel Suite, a scheduler which manages high-performance computing to reduce execution time by parallelizing the tasks.

3. Results

The workflow starts with preprocessing of the raw matrix, which consists in clustering of samples (to detect contamination), downsizing, RPKM-normalization, and metagenomics species (MGS) matrix generation. Then statistical computations between different classes of samples are performed, including richness comparisons or MGS contrast studies based on discriminant genes. The results are finally displayed as boxplots or barcodes. Furthermore, ShRCAn is designed in such a way that it is always possible to easily add new analytical steps at the end of the workflow in order to improve its capabilities.

4. Conclusion

ShRCAn is an interactive shiny application that provides a user-friendly and standardized way to process gene abundance table, from pre-processing to statistical analysis. Automation and speedup of the data preprocessing allows the bioanalyst to have more time in the interpretation of biological results.

References

[1] Edi Prifti, Emmanuelle Le Chatelier. MetaOMineR: A Quantitative Metagenomics Data Analyses Pipeline 1.1, 2014. Available from https://cran.r-project.org/web/packages/momr/index.html.
[2] Emmanuelle Le Chatelier, Trine Nielsen, Junjie Qin, Edi Prifti, Falk Hildebrand, Gwen Falony *et al*. Richness of human gut microbiome correlates with metabolic markers. *Nature* (500):541–546, 2013.
[3] H Bjoern Nielsen, Mathieu Almeida, Agnieszka Sierakowska Juncker, Simon Rasmussen, Junhua Li, Shinichi Sunagawa *et al*. Identification and assembly of genomes and genetic elements in complex metagenomic samples without using reference genomes. *Nature Biotechnology* (32): 822–828, 2014.

Genetic diversity of *Anaplasma phagocytophilum* among ticks and roe deer in a fragmented agricultural landscape

Amélie CHASTAGNER[1, 2], Angélique PION[2], Hélène VERHEYDEN[3], Bruno LOUTRET[3], Bruno CARGNELUTTI[3], Denis PICOT[3], Valérie Poux[2], Émilie BARD[2], Olivier PLANTARD[4], Karen D. MCCOY[5], Agnès LEBLOND[6], Gwenaël VOURC'H[2], and Xavier BAILLY[2]

[1] Evolutionary Ecology Group, Univ. Antwerpen, 2610 Wilrijk, Belgium

[2] EPIA, INRA, VetAgro Sup, 63122, Saint Genès Champanelle, France

[3] CEFS, UR0035, Comportement et Ecologie de la Faune Sauvage, Université de Toulouse, INRA,24 chemin de Borde-Rouge, F-31326, Castanet-Tolosan cedex, France

[4] BIOEPAR, INRA, Oniris, 44307, Nantes, France

[5] MIVEGEC (UMR 5290), Maladie Infectieuses et Vecteurs: Ecologie, Génétique Evolution et Contrôle, Centre National de la Recherche Scientifique-Université de Montpellier-Institut de Recherche pour le Développement (UR224, 911 Avenue d'Agropolis, BP 64501, F-34394 cedex 5, Montpellier, France

[6] EPIA, INRA, and Equine Medicine, VetAgro Sup, 69280, Marcy-L'étoile, France

Corresponding Author: xavier.bailly@inra.fr

Anaplasma phagocytophilum is a tick-borne pathogen infecting multiple vertebrate host species, including humans. The tick species *Ixodes ricinus* is considered as its main vector in Europe. While the prevalence of *A. phagocytophilum* in ticks is often low, *i.e.* below 5 % in questing *I. ricinus* nymphs, it can reach higher values in host populations. As an illustration, roe deer frequently shows a prevalence of *A. phagocytophilum* infection above 80 %. Roe deer are considered as a major reservoir for this pathogen, but the genotypes they carry are different from the ones that infect domestic animals and humans. We thus investigated whether roe deer was the main source of *A. phagocytophilum* genotypes circulating in questing *I. ricinus* nymphs collected either in pastures or in forest of a French fragmented agricultural landscape.

We first identified infected samples from 1,837 *I. ricinus* nymphs, sampled on geo-referenced transect lines, and 75 roe deer, tracked with GPS devices. Molecular characterization has been performed with high-throughput sequencing to take into account potential co-infections. A target region of around 350 bp was amplified by nested PCR for each marker gene: *groEL*, an housekeeping gene (coding for a chaperone protein), *msp4* and *ankA*, 2 genes associated to host specificity. We defined different alleles for each locus and analyzed the observed multilocus genotype structure.

The analysis of high-throughput *A. phagocytophilum* sequences resulted in the delineation of several genotypes for each locus: 4 for *msp4*, 5 for *ankA* and 9 for *groEL* with a frequent occurrence of co-infections. A graph approach focusing on the distribution of alleles among hosts and vectors identified two groups of alleles at the different loci, *i.e.* alleles in linkage disequilibrium, respectively associated to roe deer and tick samples. A multidimensional scaling (MD) approach combined with a multivariate analysis of variance also indicated that ticks and roe deer carried different genotypes.

A transcriptional study of five fungal *Mucor* strains

Annie LEBRETON[1], Laurence MESLET-CLADIERE[1], Jean-Luc JANY[1], Georges BARBIER[1] and Erwan CORRE[2]

[1] Laboratoire Universitaire de Biodiversité et Ecologie Microbienne (LUBEM) – Université de Bretagne Occidentale (UBO), ESIAB - Parvis Blaise Pascal - Technopôle Brest-Iroise, 29280, Plouzané, France

[2] Station biologique de Roscoff - plateforme ABiMS - CNRS : FR2424 – Université Pierre et Marie Curie (UPMC) - Paris VI, Place Georges Teissier, BP 74 29682, Roscoff CEDEX, France

Corresponding Author: corre@sb-roscoff.fr

The fungal genus *Mucor* belongs to the Mucoromycota phylum, one of the five groups of the early diverging fungi. *Mucor* species are ubiquitous, they show diverse lifestyles and may have contrasting impacts on human activities. Indeed, some pathogenic *Mucor* species represent a threat for human health, some others can be involved in food spoilage whereas some few others can be used for Asian fermented food manufacturing or cheese ripening. Despite these impacts on human activities, little is known on the genus *Mucor* and most of the studies focused only on human pathogens. Here, we are investigating on specificities linked to *Mucor* species lifestyles. We engaged a transcriptomic analysis focused on five species: *M. fuscus* and *M. lanceolatus*, two technological species used in cheese ripening, *M. racemosus*, a recurrent cheese spoiler, *M. circinelloides*, a pathogenic species and *M. endophyticus,* a plant endophyte species.

Strains of *M. fuscus* UBOCC 1.09.160 (MF), *M. lanceolatus* UBOCC 1.09.153 (ML), *M. racemosus* UBOCC 1.09.155 (MR), obtained from the *Université de Bretagne Occidentale Culture Collection*, *M. endophyticus* CBS 385-95 (ME) and *M. circinelloides* CBS 277.49 (MC) ordered from the Centraal Bureau Voor Schimmelculture CBS were grown on a standard medium and total RNA was extracted. Between 25 millions (ML) and 35 millions (ME) pairs of reads were obtained from the paired end sequencing. Transcriptomes were assembled *de novo* with Trinity, low coverage transcripts were detected with RSEM and removed from transcriptomes. Since the percent of gene with isoforms was different among species (from 7% for ME to 31% for ML), a single transcript per Trinity gene was selected to create the studied transcriptomes. Completeness of these new transcriptomes were assessed with BUSCO. Ribosomal RNA were detected with RNAmmer, CDS were predicted with Transdecoder. On these predicted proteomes, protein domains, signal peptides and transmembrane domains were annotated with respectively HmmScan against PFAM-A, signalP and tmhmm. Homologies were searched against Uniref90, swissprot-uniprot and tree Mucoromycota species. To this functional annotation were added GO terms and EC numbers transferred from Mucoromycota homologies. EC numbers were also inferred by profile detection with PRIAM. Orthogroups were predicted with OrthoFinder allowing us to propose a core transcriptome and transcripts specific to groups of species. Each of these groups of transcripts was functionally characterized by examining the composition of functional annotation (GO terms, EC numbers, protein signal...). A special interest was given for groups composed only by transcripts of species sharing the same lifestyle such as the technological species MF and ML. In the same time was investigated the repartition of genes suspected to show specificities linked to *Mucor* species lifestyles like excreted proteins and secondary metabolites. These investigations allowed to spot groups of transcripts that could be linked to a given species. This study is expected to give new hints regarding adaptation to specific habitat and/or lifestyle.

Multi-Cloud deployment for microbial genomes analysis

Jonathan LORENZO[1], Bryan BRANCOTTE[1], Thomas LACROIX[2] , Mohamed BEDRI[1], Jean-François GIBRAT[1] and Christophe BLANCHET[1]

[1] CNRS, UMS 3601 ; Institut Français de Bioinformatique, IFB-core, Avenue de la Terrasse, F-91190 Gif-sur-Yvette, France

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

Corresponding Author: jonathan.lorenzo@france-bioinformatique.fr, Thomas.lacroix@inra.fr

In the post-NGS area, sequencing bacterial genomes is very cheap (few hundreds €). Most of the time, users are no longer content to analyse a single genome; they want to compare large collections of related genomes (strains). This entails that biologists have to pay too much attention and dedicate their time to sequence the genomes, instead of thoroughly analysing the genomic data. Thus, this brings light to the increasing need for automating the annotation of bacterial genomes and carrying out efficient data mining.

In that context, IFB hub and the IFB-MIGALE platform developed a virtual environment (appliance), based on virtual machines, called "bacterial genomics" that aims to provide biologists and bioinformaticians access to suitable resources via the cloud. For example, Prokka [1] is a software tool for the rapid annotation of prokaryotic genomes. Insyght [2] developed by IFB-MIGALE is a tool for the visualization of the syntenies (local conservation of the gene order along the genomes) and the exploration of the landscape of both conserved and idiosyncratic genomic regions across multiple genomes. The platform automatically launches a set of bioinformatics tools (e.g. BLAST, HMMER, Prodigal…) to analyse the data and stores the results in a relational database (PostgreSQL). These tools use several public reference data collections. A web interface allows the user to browse the results. Setting up the platform requires solid skills in system administration since many bioinformatics tools with different dependences need to be installed as well as a relational database management system, a web server and servlet container, etc. Moreover, performing the analysis of a large number of genomes requires large computing resources and the use of parallel computing.

The goal is to deploy the “appliance” in one click over one or more cloud infrastructures. To achieve this, new features to automate deployment of complex application were added to the IFB’s cloud portal [3] through the connection to the SlipStream cloud broker [4]. Developed by SixSq, SlipStream is a multi-cloud application management platform. It automates the full application management lifecycle, within Infrastructure as a Service (IaaS) cloud infrastructures. Such complex application deployments can be done over several cloud infrastructures and provide scientists with high-level cloud features such as the dynamic allocation of a dedicated network for the isolation of the virtual machines, with the replication of the user data and with a direct link from the cloud portal to the Insyght web portal [5]. The appliance is available in the RAINBio catalogue of virtual images on the Biosphere web portal [6], and several tutorials on IFB bioinformatics cloud services usage are also available online on the main IFB website.

[1] Seemann T. Prokka: rapid prokaryotic genome annotation. Bioinformatics. 2014 Jul 15;30(14):2068-9. PMID:24642063

[2] Lacroix T., Loux V., Gendrault A., Hoebeke M., and Gibrat J.F. Insyght: navigating amongst abundant homologues, syntenies and gene functional annotations in bacteria, it's that symbol! Nucleic Acids Res. 2014 Dec 1; 42(21): e162. doi: 10.1093/nar/gku867 ; PMCID: PMC4245967

[3] https://biosphere.france-bioinformatique.fr/

[4] http://sixsq.com/products/slipstream/index.html

[5] https://cyclone.france-bioinformatique.fr/usecases/view/125

[6] https://biosphere.france-bioinformatique.fr/catalogue/appliance/19/

Evaluation of 15 in silico prediction tools for the classification of *MED13L* missense variations

Thomas SMOL[1,2], Caroline THUILLIER[2], Sylvie MANOUVRIER-HANU[1,3] and Jamal GHOUMID[1,3]

[1] RADEME Research Team EA7364, Université de Lille, 59000, Lille, France
[2] Institut de génétique médicale, CHRU Lille, 59000, Lille, France
[3] Service de génétique clinique, CHRU Lille, 59000, Lille, France

Corresponding Author: thomas.smol@chru-lille.fr

Abstract

Introduction. *MED13L* is a well conserved gene involved in the structure of the Mediator Complex. Truncating variations were reported in patients with intellectual disability (ID). Hence, *MED13L* was included in sequencing panels for ID diagnosis and the analysis of missense variations was not obvious. Here, we compared 15 bioinformatics predictive tools to aid in the *MED13L* missense interpretations.

Methods. All missense positions for *MED13L* were extracted from ExAC database, and considered as class-1 or 2 [1] if there were also present in EVS database or if the allele counts were upper than 1 in ExAC. All class-4 and 5 missenses [1] were found in Clinvar, Decipher, Pubmed, or in our in-house *MED13L* database as french referent laboratory for *MED13L* pathogenic variations for the healthcare network "DefiScience". Fifteen *in silico* predictive tools were used. They were based on evolutionary conservation (FATHMM, MutationAssessor, PhD-SNP, PANTHER, SIFT), on protein structure/function and evolutionary conservation (Align GVGD, MutationTaster, PolyPhen2, SNAP2, UMD-predictor), on protein function (SNPs&GO), on similarity between variant sequence and homolog sequence (PROVEAN), or on combined tools (CONDEL, CADD and REVEL) with different cut-offs. We compared sensibility (SE), specificity (SP), accuracy (ACC), and Matthew's correlation coefficients (MCC) [2]. A second analysis was performed on missense considered as deleterious for CONDEL. All databases were consulted in January 2017.

Results. We analysed 513 missense variations from *MED13L*. Eleven missenses were considered as class-4 or 5 and 502 as class-1 or 2. All pathogenic missenses were never found in ExAC database. Their median PhyloP score was 5.731 [4.304 to 6.172] and median Grantham score was 94 [15 to 145]. The median number of allele count in ExAC for benign missenses was 7, with a mean count of 25, a median PhyloP score of 2.872 [-2.455 to 6.318] and a median Grantham score of 58 [0 to 215]. CONDEL showed the best performance in MCC measures with 0.33, and all pathogenic variations were properly classified (SE = 100%). With caution, due to the small number of pathogenic missenses, the higher SE values (100%) were obtained with CONDEL, Polyphen2, CADD (cut-off 15 and 20), UMD-predictor, PANTHER and MutationTaster. Conversely, these 6 tools shared lower SP. After filtering on non-deleterious variations for CONDEL (n = 239), CADD (cut-off 20) showed the best compromise with a SE of 100%, a SP of 22.67%, and a MCC measure of 0.12.

Conclusions. Considering MCC as the best parameter to measure predictor's performance and the need to keep all pathogenic variations, CONDEL seems to present the best compromise. Moreover, combination between combined tools CONDEL and CADD (cut-off 20) has allowed us to reduce the number of class 1 and 2 variations without affecting the number of class-4 and 5 variations in *MED13L* gene analysis.

References

[1] Richards S, *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.

[2] Johnson MM, *et al*. *Screening for deleterious nonsynonymous single-nucleotide polymorphisms in genes involved in steroid hormone metabolism and response*. Cancer Epidemiol Biomark. 2005 May;14(5):1326–9.

HiFit: robust data analysis method for High-throughput qPCR

Mathieu BAHIN[1], Quentin VIAUTOUR[1,2], Elise DIAZ[2,3], Bertrand Ducos[2,3,4] and Auguste GENOVESIO[1]

[1] Scientific Center for Computational Biology, Institut de Biologie de l'ENS, 46, rue d'Ulm, 75005, Paris, France

[2] High Throughput qPCR Facility, Institut de Biologie de l'ENS, 46, rue d'Ulm, 75005, Paris, France

[3] LPS-ENS, CNRS UMR8550, 46, rue d'Ulm, 75005, Paris, France

[4] Laser Microdissection Facility, CIRB Collège de France, 11 Place Marcelin Berthelot, 75005, Paris, France

Corresponding Author: mathieu.bahin@biologie.ens.fr

Real-time quantitative polymerase-chain-reaction (RT-qPCR) is frequently used as a standard technique for various applications such as research or clinical diagnostic. To date, several data analysis strategies have been proposed to extract meaningful information from single RT-qPCR curves [1,2,3]. Most of them are in fact semi-automated because they were developed in the context of low-throughput RT-qPCR data where each reaction can be visually investigated and its analysis manually corrected. We observed that portability of those methods to high throughput data was far to be straightforward and lacked robustness. In fact several of them could simply not be fully automated. To address this question we have developed an absolute high throughput qPCR data analysis approach based on a robust fitting of a four or six parameters sigmoid model. We take advantage of the throughput such that the search of the optimal parameters for each curve is achieved using information gathered from the fitting of large sets of curves obtained from the whole dataset. Our approach brings the level of robustness required to address high throughput qPCR data.

References

[1] Kenneth J Livak and Thomas D Schmittgen. Analysis of relative gene expression data using real-time quantitative pcr and the 2-δδct method. *Methods*, 25(4):402-408, 2001.

[2] DV Rebrikov and D Yu Trofimov. Real-time pcr: a review of approaches to data analysis. *Applied Biochemistry and Microbiology*, 42(5):455-463, 2006.

[3] Robert G Rutledge and Don Stewart. Assessing the performance capabilities of lre-based assays for absolute quantitative real-time pcr. *PLoS One*, 5(3):e9731, 2010.

Is UniProtKB Missing Knowledgeable Proteins?

Benoit BELY[1], Sara BENMOHAMMED[1], Guoying QI[1], Nidhi TYAGI[1] and Maria MARTIN[1]

[1] Protein Function Development EMBL-EBI, Wellcome Trust Genome Campus, CB10 1SD, Hinxton, UK

Corresponding Author: benoit.bely@ebi.ac.uk

1 Introduction

In 2015 UniProt removed 46.9 million unreviewed UniProtKB/TrEMBL records found in redundant proteomes (see: http://www.uniprot.org/help/proteome_redundancy). The redundancy pipeline [1] computes proteome redundancy for bacteria proteomes and, since UniProt 2016_08 release, for fungi proteomes as these two divisions constitute the majority of the new proteomes sequenced. Since 2015 the number of removed records has steadily increased such that UniProtKB (2017_01 release) was composed of 81.2 million protein identifiers (PIDs) from the INSDC (International Nucleotide Sequence Database Collaboration), generating 73.7 million UniProtKB/TrEMBL records with 153.5 million PIDs excluded because they were classified as redundant. Therefore, two-thirds (65.3%) of the INSDC proteins are missing in UniProtKB and by removing these sequences is the UniProtKB no longer a comprehensive resource?

2 153 million protein identifiers not in UniProtKB

A large proportion of new genomic sequencing data are from closely related organisms already sequenced and published in the INSDC. This means that the protein sequences from these organisms are not unique but do generate new PIDs. Therefore, these redundant protein sequences are stored in UniParc, UniProt's sequence archive by default. It contains all the protein sequences, including those from the INSDC, where a unique sequence will have a unique UniParc identifier (UPI) and then identifier cross-references to database sources having the unique sequence. Therefore, a UPI can have multiple links to UniProtKB and INSDC as long as the sequence represented by that UPI is the same. Among the 153.5 million PIDs not integrated into UniProtKB 149.7 million have a UPI which also links to a UniProtKB accession. Only 3.8 million PIDs (2.5%) represented in 666,988 UPIs (unique sequences) do not have identical sequence in UniProtKB.

Question: Of the 666,988 missing sequences is UniprotKB missing important functional knowledge that requires new UniProtKB/TrEMBL records with unique annotations? Before answering this question, it is important to know that annotations made in UniProtKB/TrEMBL come from automatic annotation pipelines using InterPro integrated signatures and if two different sequences contain the same set of InterPro integrated signatures, then both records will have exactly the same annotation. We evaluated if any of the 666,988 UPIs match a sequence in the Pan-Proteomes or UniRef90 using Blastp. Then we analyzed whether statistically significant subject/query hits have a differential InterPro integrated signature set by using InterProScan to see if UPIs would make new unique automatic annotation. 65,238 and 582,164 UPIs have a significant hit (identity>80% and hit coverage>50% length of query and subject) in Pan-Proteomes and UniRef90, respectively. From these hits, 122 UPIs have a differential InterPro integrated signature set that could potentially lead to a new annotation. However, 18,803 UPIs did not return a significant Blastp hit; where 9,811 UPIs are fragments, identity and coverage of the query is > 90% but subject coverage is < 50%. The other 8,992 UPIs, 2,643 UPIs have InterPro signature set that could potentially lead to a new annotation.

3 Conclusion

It is correct not to integrate 65.3% of INSDC into UniProtKB. The CDSs from INSDC have a high level of redundancy, 97.5%, to CDSs sequences already present in UniProtKB. Of the remaining 2.5% sequences that are technically missing from the UniProtKB only 4.2% of these sequences (2,765/666,988) could potentially add valuable knowledge to the knowledgebase by inferring new functional annotations.

References

[1] Bursteinas B, Britto R, Bely B, Auchincloss A, Rivoire C, Redaschi N, O'Donovan C and Martin MJ. Minimizing proteome redundancy in the UniProt Knowledgebase *Database (Oxford)*, 2016 Dec 26, 2016.

Co-option of complex molecular system in bacterial membranes

Rémi DENISE[1,2], Sophie ABBY[3,4] and Eduardo PC ROCHA[1,2]

[1] Institut Pasteur, Microbial Evolutionary Genomics Group, 28 rue du Docteur Roux, 75015, Paris, France

[2] CNRS, UMR 3525, 28 rue du Docteur Roux, Paris, France

[3] Univ. Grenoble Alpes, Laboratoire Techniques de l'Ingénierie Médicale et de la Complexité - Informatique, Mathématiques et Applications, Grenoble (TIMC-IMAG), F-38000 Grenoble, France

[4] CNRS, TIMC-IMAG, F-38000 Grenoble, France

Corresponding Author: remi.denise@pasteur.fr

Protein secretion systems exist in many bacterial and archaeal species and are important for bacterial virulence. These systems are complex machineries made of many different proteins that interact together to allow other proteins to pass through the cell wall and be secreted outside of the cell. The proteins constituting these systems show various evolutionary rates and patterns of conservation within the secretion systems. Many of these systems were co-opted in complex evolutionary processes from other molecular structures.

The co-option is the emergence of new complex forms and molecular systems from other systems with different functions. This process of co-option (or exaptation) of functional or structural traits is thought to be a major driver of functional innovation [1]. The appendages of bacteria provide some striking examples of these processes. For example, the type III secretion system (T3SS) was co-opted from the bacterial flagellum [2].

Biochemical, phylogenetic, and structural evidence show that the family of molecular machineries including type II secretion system (T2SS, involved in protein secretion), type IV pilus (T4P, involved in cell motility, adherence and virulence), Tad pilus (idem), the competence apparatus (Com, involved in natural transformation) and the archaeal flagellum (Archaellum, motility) share homologous genes and have a similar genetic organization [3].

We designed custom comparative genomic tools to detect and distinguish macromolecular systems in genome sequences, based on their particular components and genetic organization (MacSyFinder, [4]). This allows us to investigate the evolutionary origins of these machineries by phylogenetic and comparative genomics approaches, and thus to decipher some mechanisms of co-option involved in the diversification of microbial cellular machineries. Another goal would be the use of these phylogenetic analyses to facilitate the discrimination between related systems, and produce tools to perform the automatic annotation of an unknown system.

We have detected more than 6400 systems of these family of systems among all the bacteria (mostly in Proteobacteria), and we have identified their key components (ATPase, inner membrane platform, major pilin, prepilin peptidase, secretin) on a dataset of more than 5750 complete genomes. For each key component, we've established a phylogeny, and we are now trying to reconcile them and understand the biological reasons of their discordance by inferring the phylogeny of this family of systems.

The systems analyzed are probably among the most complex network of molecular co-options analyzed to date and should provide an excellent basis to (i) infer the frequency of horizontal transfer of each type of derived molecular system, and (ii) study the evolution of the genetic organization of the loci encoding these systems in the light of their evolutionary history.

1. Gould, S.J. and E.S. Vrba, *Exaptation-A Missing Term in the Science of Form.* Paleobiology, 1982. **8**: p. 4-15.
2. Abby, S.S. and E.P. Rocha, *The non-flagellar type III secretion system evolved from the bacterial flagellum and diversified into host-cell adapted systems.* PLoS Genetics, 2012. **8**(9): p. e1002983.
3. Korotkov, K.V., M. Sandkvist, and W.G. Hol, *The type II secretion system: biogenesis, molecular architecture and mechanism.* Nature Reviews. Microbiology, 2012. **10**(5): p. 336-51.
4. Abby, S.S., et al., *MacSyFinder: A Program to Mine Genomes for Molecular Systems with an Application to CRISPR-Cas Systems.* PLoS ONE, 2014. **9**(10): p. e110726.

High-quality, fast, and memory-efficient assembly of metagenomes and large genomes using Minia-pipeline

Rayan Chikhi[1], Charles Deltel[2], Guillaume Rizk[2], Claire Lemaitre[2], Pierre Peterlongo[2], Kristoffer Sahlin[4], Lars Arvestad[3], Paul Medvedev[4]and Dominique Lavenier[2]
[1] CRIStAL, (UMR CNRS 9189), France.[2] IRISA, Univ Rennes, France.[3] Stockholm University, Sweden.[4] Pennsylvania State University, USA.

Corresponding author: rayan.chikhi@univ-lille1.fr

1 Abstract

Gigabase-scale genome projects and large metagenomics studies have flourished thank to high-throughput sequencing technologies. However, performing de novo assembly of such data remains challenging. In the landscape of assembly software, the tools that produce high-quality assemblies typically require significant computational resources, while the fast and memory-efficient ones yield relatively inferior results. We present Minia-pipeline: an assembler that combines efficiency and high-quality results. Minia-pipeline is geared towards large datasets of metagenomes and eukaryotic genomes, and recently provided high-ranking assemblies in the Critical Assessment of Metagenomic Interpretation challenge. This poster describes the overall architecture of the pipeline, key algorithmic improvements, and demonstrate its effectiveness on both large genome and metagenome samples. The pipeline is modular and integrates several components: an error-correction module, a unitig assembly tool (BCALM 2 [1]), a multi-k contigs assembly module (Minia 3), and a scaffolder (BESST [2]). Software is available at https://github.com/GATB/gatb-minia-pipeline

References

[1] Rayan Chikhi, Antoine Limasset, and Paul Medvedev. Compacting de bruijn graphs from sequencing data quickly and in low memory. *Bioinformatics*, 32(12):i201–i208, 2016.

[2] Kristoffer Sahlin, Rayan Chikhi, and Lars Arvestad. Assembly scaffolding with pe-contaminated mate-pair libraries. *Bioinformatics*, page btw064, 2016.

Debugging long-read genome and metagenome assemblies using string graph analysis

Pierre MARIJON[1], Jean Stéphane VARRÉ[2] and Rayan CHIKHI[2]
[1] Inria, Université de Lille, CNRS, Centrale Lille, UMR 9189 - CRIStAL - Centre de Recherche en Informatique Signal et Automatique de Lille, F-59000 Lille, France
[2] Univ. Lille, CNRS, Centrale Lille, Inria, UMR 9189 - CRIStAL - Centre de Recherche en Informatique Signal et Automatique de Lille, F-59000 Lille, France

Corresponding author: pierre.marijon@inria.fr

Third-generation long-read sequencing technologies tackle the repeat problem in genome assembly by producing reads that are long enough to span most repeat instances. In principle one expects that with such reads most bacterial genomes will be assembled into a single contig [1]. However in practice, some datasets fail to be perfectly assembled even with leading assemblers, and are fragmented into a handful of contigs. As a mean to investigate those cases, we consider the string graphs that are generated by assemblers during intermediate stages of the assembly process. We seek to establish a coherent framework for analyzing these graphs, in the hope that they will help us determine the biological causes that led the assembler to output shorter contigs. This poster presents some preliminary results of such an analysis.

We visualized, analyzed and compared assembly graphs generated by *Canu* [2] and *Miniasm* assemblers [3] on biological (MBRAC-26 [4]) and synthetic datasets (created with LongISLND [5]). We introduce the concept of *graph projection* of an assembly graph onto another, taking advantage of the recent GFA format. We are thus able to observe how reads that are neighbors of contigs extremities overlap, in terms of error rate and overlap length. We implemented an automatic and user-friendly *snakemake* pipeline that generates a HTML report for each assembly. We identified cases of contigs that were not joined by the assembler despite indications in the string graph that such joins could have been made. These cases highlight potential directions on how to improve the assembly process. In future work we will take advantage of this investigation to propose alternative assembly hypotheses based on string graph analysis.

References

[1] Sergey Koren and Adam M Phillippy. One chromosome, one contig: complete microbial genomes from long-read sequencing and assembly. *Current Opinion in Microbiology*, 23:110–120, 2015.

[2] 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*, page gr.215087.116, 2017.

[3] Heng Li. Minimap and miniasm: fast mapping and de novo assembly for noisy long sequences. *Bioinformatics*, 32(14):2103–2110, 2016.

[4] Esther Singer, Bill Andreopoulos, Robert M. Bowers, Janey Lee, Shweta Deshpande, Jennifer Chiniquy, Doina Ciobanu, Hans-Peter Klenk, Matthew Zane, Christopher Daum, Alicia Clum, Jan-Fang Cheng, Alex Copeland, and Tanja Woyke. Next generation sequencing data of a defined microbial mock community. *Scientific Data*, 3:160081, 2016.

[5] Bayo Lau, Marghoob Mohiyuddin, John C. Mu, Li Tai Fang, Narges Bani Asadi, Carolina Dallett, and Hugo Y. K. Lam. LongISLND: in silico sequencing of lengthy and noisy datatypes. *Bioinformatics*, 32(24):3829–3832, 2016.

Labsquare une communauté de développeurs libres http://www.labsquare.org

Sacha SCHUTZ[1], Pierre MARIJON[1], Jérémie ROQUET[1], Francisco PINA-MARTINS[1], June SALLOU[1], Eugene TROUNEV[1], Anne-Sophie DENOMMÉ[1] and Olivier GUEUDELOT[1]
Labsquare Team

Corresponding author: sacha@labsquare.org

1 Objectif

Avec l'évolution des technologies de séquençage haut débit, la bioinformatique est devenue un point de passage obligatoire en biologie ou encore en médecine pour le diagnostic moléculaire. L'analyse de ces données génomiques nécessite le plus souvent des compétences informatiques hors de portée d'un non-bioinformaticien qui souhaite gagner en autonomie. Plusieurs startups ont d'ores et déjà pris le devant pour répondre à cette demande en produisant des logiciels payants accessibles à tous à l'aide d'interfaces graphiques simples et épurées. Labsquare est une communauté de développeurs qui veut se poser comme alternative aux solutions commerciales en produisant des interfaces graphiques libres et accessibles dans le domaine de la génomique en suivant le même modèle que certaines associations à but non lucratif comme Framasoft ou KDE.

2 Communauté

Nous sommes pour l'instant une petite équipe composée de bioinformaticiens, developpeurs, designers, médecins et généticiens. Nous communiquons et travaillons ensemble grâce à des outils comme Github, Gitter ou Framatalk.

3 Technologie

Le framework C++ Qt est une de nos technologies préférées en nous permettant de réaliser des interfaces graphiques modernes et multiplateformes. A titre d'exemple Rstudio, Mendeley, Bandage et Alamut sont codé avec Qt. Les applications labsquare sont et seront toujours libres sous licence GPL3 en respectant les standards du GA4GH (Global Aliance for Genomics and Health).

4 Applications

Quatres applications sont déjà disponibles ou en cours de developpement.

FastQt est le clone de FastQC[1] et permet d'analyser des fichiers FASTQ.
https://github.com/labsquare/fastQt

CuteVCF est un viewer de fichier VCF.
La prochaine version, CuteVariant se calquera sur le fonctionnement de variant-tools[2].
https://github.com/labsquare/CuteVCF

CutePeaks Un simple visualisateur de fichier AB1 (Sanger trace file) .
https://github.com/labsquare/CutePeaks

BigBrowser Un genome browser mixant les caractéristiques d'IGV[3] [4] et d'Alamut visual.
https://www.youtube.com/watch?v=Y7ouuS8Ooo0

References

[1] Simon Andrews et al. Fastqc: a quality control tool for high throughput sequence data, 2010.

[2] San Lucas, F Anthony, Gao Wang, Paul Scheet, and Bo Peng. Integrated annotation and analysis of genetic variants from next-generation sequencing studies with variant tools. *Bioinformatics*, 28(3):421–422, 2012.

[3] Helga Thorvaldsdóttir, James T Robinson, and Jill P Mesirov. Integrative genomics viewer (igv): high-performance genomics data visualization and exploration. *Briefings in bioinformatics*, 14(2):178–192, 2013.

[4] James T Robinson, Helga Thorvaldsdóttir, Wendy Winckler, Mitchell Guttman, Eric S Lander, Gad Getz, and Jill P Mesirov. Integrative genomics viewer. *Nature biotechnology*, 29(1):24–26, 2011.

Influence of SNP coding on the analysis of disease risk

Hélène SARTER[1], Corinne GOWER-ROUSSEAU[1], Guillemette MAROT[2,3]

[1] Lille Inflammation Research International Center – U995, F-59000 Lille, France
[2] Univ. Lille, EA 2694 - Santé publique : épidémiologie et qualité des soins, F-59000 Lille, France
[3] Inria Lille Nord Europe, MODAL, F-59000 Lille, France

Corresponding Author: helene.sarter@chru-lille.Fr

1 Introduction

In genome wide association studies, although it is a strong biological assumption, SNP (Single Nucleotide Polymorphisms) additive model is often used before testing the association between a phenotype of interest and SNPs. SNP additive model refers to the coding of SNP by the number of rare allele and allows to consider SNPs as continuous variables and to use statistical methods that only allow continuous predictors (for example lasso or PLS regression). Yet, data on the impact of this choice is lacking. The objective of our study was to test, in a simulation framework, the influence of SNP coding on the selection of SNPs linked to a phenotype of interest.

2 Methods

We used data from EPIMAD registry [1] on the severity of Crohn's disease. 156 patients have been genotyped for 369 variants. We used the following simulation framework : the 12 most significant SNPs in univariate analysis were considered as "influent" and others as "non influent". We randomly permutated all "non influent" SNPs, using the same permutation for all SNPs in order to keep the correlation structure between SNPs. "Influent" genes were not permuted in order to keep their relation with the dependent variable. This sample set was used to test the selection of SNPs with the lasso [2] and stability selection method [3] with 5 SNPs codings : additive, dominant, recessive and heterozygote models and finally the group-lasso method [4] that allows to code SNPs as dummy variables.

3 Results

With a threshold of 0.6 in stability selection, heterozygote and group-lasso models permitted to select more "influent" SNPs than other models, including the additive model : heterozygote and group-lasso selected 7/12 SNPs whereas additive model selected only 3/12. All models failed to select all "influent" SNPs and to avoid false positives. This might be related to the simulation framework using real data : "influent" SNPs might be only slightly related to the dependent variable or correlated to other SNPs. Our study highlighted that the additive genetic model can fail to select the real variables and that more interest should be given to the research and use of statistical models allowing qualitative variables, especially group methods in the field of sparse regression methods. Yet, our study needs to be replicated on another public data set with more strong association of SNPs and the dependent variable. Sparse-PLS regression methods need also to be tested since they might better take correlation between SNPs into account.

References

[1] Gower-Rousseau C., Vasseur F., Fumery M., Savoye G., Salleron J., Dauchet L., Turck D., Cortot A., Peyrin-Biroulet L., Colombel J.F. Epidemiology of inflammatory bowel diseases: new insights from a French population-based registry (EPIMAD). Dig Liver Dis. 2013 Feb;45(2):89-94.
[2] Tibshirani R. (1996). Regression shrinkage and selection via the lasso. *J. Royal. Statist. Soc B.*, Vol. 58, No. 1, pages 267-288.
[3] Meinshausen N. and Bühlmann P. (2010). Stability selection (with discussion). *J. Royal. Statist. Soc B*, 72, 417-473.
[4] Yuan M. Model selection and estimation in regression with grouped variables. *J. Royal. Statist. Soc B.*, Vol. 68, No. 1, pages 49-67.

Ultra High throughput, single molecule mapping of replicating DNA

Nikita MENEZES BRAGANCA[1], Francesco DE CARLI[1], Wahiba BERRABAH[2], Valérie BARBE[2], Auguste GENOVESIO[1] and Olivier HYRIEN[1]

[1] Institut de Biologie de l'Ecole Normale Supérieure (IBENS),46 rue d'Ulm, 75005 Paris, France

[2] Génoscope, 2 rue Gaston Crémieux, 91057 Evry, France

Corresponding Author: menezes@biologie.ens.fr

Visualization and mapping of DNA molecules at single molecule level genome-wide is possible thanks to a new device based on microfluidics developed by BionanoGenomics. DNA fibers stained with a fluorescent intercalator, YOYO-1, and labeled by incorporation of fluorescent nucleotides using a nicking endonuclease creating a restriction map (barcode), are stretched in nanochannel arrays and imaged automatically. We typically collect, in a single run, over 32 000 images and more than 15 000 Mbp of DNA in the form of >100 Kbp long DNA molecules. This device, made for finding structural variations in a given genome and for *de novo* assembly, is now being used for the study of replication origins.

From each of the images the DNA molecule intensity profile is extracted after having achieved preprocessing and registration steps on the raw images. Based on the analysis of these one dimension profiles, we would like to know if there is a correlation between tracts of higher intensity (doubling) and replication bubbles that has been observed in previous studies performed in our lab. To do so, adaptation of the provided proprietary softwares and new tool development are required. One of the tools that I have already implemented, enables us to visually check the DNA molecule detection and the optical mapping performed and see where origins of replication are fired molecule by molecule genome wide. I will be presenting the current effort I put to automatically analyze the thousands of DNA molecules in order to validate our observations regarding this doubling of intensity.

References

[1] Hyrien, O. Peaks cloaked in the mist: The landscape of mammalian replication origins. *Journal of Cell Biology*, **208**, 147–160. doi:10.1083/jcb.201407004, 2015

[2] Hyrien, O. *The Initiation of DNA Replication in Eukaryotes*, D Kaplan, Ch. 4, Springer, 2016

[3] Hyrien, O. *and al.* From simple bacterial and archaeal replicons to replication N/U-domains. *Journal of Molecular Biology*. doi:10.1016/j.jmb.2013.09.021, 2013

[4] Michalet, X. Dynamic Molecular Combing: Stretching the Whole Human Genome for High-Resolution Studies. *Science*, *277*(5331), 1518–1523. doi:10.1126/science.277.5331.1518, 1997

[5] De Carli, F., Gaggioli, V., Millot, G. A., & Hyrien, O. Single-molecule, antibody-free fluorescent visualisation of replication tracts along barcoded DNA molecules. *The International Journal of Developmental Biology*, (May), 1–9. doi:10.1387/ijdb.160139oh, 2015

Data updates on Norine, the reference Non-Ribosomal Peptide knowledge base

Yoann DUFRESNE[1], Juraj MICHALIK[1], Areski FLISSI[1], Valerie LECLÈRE[1,2] and Maude PUPIN[1]
[1] Équipe Bonsai, Univ. Lille, CNRS, Centrale Lille, UMR 9189 - CRIStAL - Centre de Recherche en Informatique Signal et Automatique de Lille, F-59000 Lille, France
[2] Équipe ProBioGEM, Univ. Lille, INRA, ISA, Univ. Artois, Univ. Littoral Côte d'Opale, EA 7394 - ICV - Institut Charles Viollette, F-59000 Lille, France

Corresponding author: yoann.dufresne0@gmail.com

1 The Norine database

Norine, first released in 2006 [1], remains the unique platform dedicated to computational analysis of non-ribosomal peptides (NRPs). Among others, NRPs can act as antibiotics, siderophores, surfactants or protease inhibitors. The Norine database is the reference NRP knowledge base, containing more than 1200 peptides composed of almost 530 different monomers (various building blocks including amino acids). Each referenced NRP have a dedicated web page with various informations, including the most important, their composition and their biological activities. The monomeric representation, correspond to the nearest representation of the NRP assembly process. The other representation is the atomic representation, obtained by reconstruction after a mass spectrum analysis. The knowledge of the monomeric representation allow to understand the synthesis pathway. It has also been proved [2] that, the activity of the molecule can be predicted from this representation.

2 Improving the data quality and quantity

We developed a tool called smiles2monomers (s2m) [3] that automatically creates NRP annotations. From a SMILES [4], s2m infers the monomeric structure of the NRP. In Norine, a significant amount of NRP entries (around 30%) are annotated with both structures. We used s2m on the atomic structures to verify the integrity of the data and we found a few errors (50 NRPs with a wrong atomic or monomeric structure). To avoid the insertion of new errors, we included the s2m software in the crowdsourcing tool MyNorine.

We identified 3 main databases that could be sources of new NRPs for Norine: MIBiG [5] (stores gene clusters of secondary metabolites), BIRD [6] (Centralisation of external resources about "interesting" molecules), StreptomeDB [7] (molecules produced by bacteria in the *Streptomyces* genus). The Norine database is well known for the quality of its manual annotations. So, we did not want to add wrong informations from an automatic filling of the database. For this reason, we created a strict validation pipeline for the potential new entries. After the filtering process, we found 472 NRPs unreferenced in Norine: 235 from MIBiG, 162 from BIRD and 75 from StreptomeDB. Those data represent an increase of 30% of the entries in the Norine database.

3 Conclusion

In this poster we present an update of the data from the knowledge base Norine. Using tools like smiles2monor we detected a few errors in the annotations. We corrected them and created safeguards to avoid errors in future user submissions. In a second time, we used several tools to retrieve and filter many possible new NRP entries in the database. The work on automatic filing scripts led us to a data increase of 30%. So, in the coming release of Norine we strongly improve the data quantity and quality available for all.

References

[1] Ségolène Caboche, Maude Pupin, Valérie Leclère, Arnaud Fontaine, Philippe Jacques, and Gregory Kucherov. NORINE: a database of nonribosomal peptides. 36:D326–D331.
[2] Ammar Abdo, Ségolène Caboche, Valérie Leclère, Philippe Jacques, and Maude Pupin. A new fingerprint to predict nonribosomal peptides activity. 26(10):1187–1194.
[3] Yoann Dufresne, Laurent Noé, Valérie Leclère, and Maude Pupin. Smiles2monomers: a link between chemical and biological structures for polymers. 7:62.
[4] David Weininger. SMILES, a chemical language and information system. 1. introduction to methodology and encoding rules. 28(1):31–36.
[5] MIBiG: Minimum information about a biosynthetic gene cluster.
[6] Helen Berman, Kim Henrick, and Haruki Nakamura. Announcing the worldwide protein data bank. 10(12):980–980.
[7] Xavier Lucas, Christian Senger, Anika Erxleben, Björn A. Grüning, Kersten Döring, Johannes Mosch, Stephan Flemming, and Stefan Günther. StreptomeDB: a resource for natural compounds isolated from streptomyces species. 41:D1130–D1136.

NRPro: a Bioinformatics Tool for Nonribosomal Peptides Identification by Tandem Mass Spectrometry

Emma RICART[1], Mickael CHEVALIER[2], Maude PUPIN[3,4], Valerie LECLERE[2], Christophe FLAHAUT[2] and Frederique LISACEK[1]

[1] Proteome Informatics Group, Swiss institute of Bioinformatics and University of Geneva, CUI-7 Route de Drize, CH-1211, Geneva, Switzerland

[2] University of Lille, INRA, ISA, University of Artois, Univ. Littoral Côte d'Opale, EA 7394-ICV- Institut Charles Viollette, F-59000 Lille, France

[3] University of Lille, CNRS, Centrale Lille, UMR 9189- CRIStAL- Centre de Recherche en Informatique Signal et Automatique de Lille, F-59000 Lille, France

[4] Inria-Lille Nord Europe, Bonsai team, F-59655 Villeneuve d'Ascq Cedex, France

Corresponding Author: emma.ricart@sib.swiss

Nonribosomal Peptides (NRPs) are natural compounds enzymatically synthetized by microorganisms such as bacteria and fungi. These peptides have shown a wide range of biological properties such as antibiotics, antitumor or immunosuppressant, being of great importance to the pharmacological and agricultural industries. Due to its high sensitivity and accuracy, Mass Spectrometry (MS) is crucial for the identification of these biomolecules. However, the unusual chemical structures of NRPs (cyclic, polycyclic, branched…) and the presence of highly modified non-proteogenic amino-acids complicate the interpretation of the MS/MS spectra. Tools for the identification of some simple NRPs already exist, but they do not cover all NRP specificities, lack flexibility, efficient scoring and statistical validation as well as user friendliness. Here we present a new bioinformatics tool to match predicted MS/MS spectra against their experimental counterparts, either exactly or in a modification tolerant way.

Norine, a database entirely dedicated to non-ribosomal peptides [1], is used to retrieve the molecular structure of the NRPs. Based on this information we have developed a fragmentation model to calculate the MS/MS fragments of each peptide and predict their theoretical spectrum. The model covers all the structures observed in NRPs (cyclic, multicyclic and branched) and includes the 500+ non-proteogenic monomers. All the known fragmentation characteristics of NRPs have been included and multiple ring breakages are taken into account in order to calculate the putative fragment masses. Additionally, a combinatorics algorithm has been developed in order to allow modification and adduct tolerant searches. Once a spectrum-peptide match (PSM) is confidently identified, it can be added to a spectral library with its corresponding annotations.

Our software is presented as a web application developed in Javascript, CSS and HTML for the client side and Java for the server side. It provides a highly interactive interface and it is able to perform a configurable and complete computational fragmentation of NRPs, including those presenting complex structures containing multicycles and several branches. Preliminary tests with experimental MS/MS data show positive results: the tool is able to match all the high intensity peaks. Furthermore, this is the first NRP fragmentation tool that includes modification tolerant searches, which will be very useful for the identification of new peptides.

Acknowledgements

We would like to thank the Swiss Institute of Bioinformatics for providing financial support for this project, Yoann Dufresne for his advises regarding the usage of his bioinformatics tool [2] and Markus Müller for sharing his expertise on mass spectrometry and MzJava [3].

References

[1] Flissi Areski, Dufresne Yoann, Michalik Juraj, Tonton Laurie, Janot Stephane, Noe Laurent, Jacques Philippe, Leclere Valerie and Pupin Maude: Norine, the knowledgebase dedicated to non-ribosomal peptides, is now open to crowdsourcing. In *Nucleic Acids Research 44(D1)*, pages D1113–D1118. Oxford Academic, 2016.

[2] Dufresne Yoann, Noe Laurent, Leclere Valerie and Pupin Maude: Smiles2Monomers: a link between chemical and biological structures for polymers. In *Journal of cheminformatics 7(1)*, page 62. Springer, 2015.

[3] Horlacher Oliver, Nikitin Frederic, Alocci Davide, Mariethoz Julien, Müller Markus and Lisacek Frederique: MzJava: An open source library for mass spectrometry data processing. In *Journal of Proteomics 129*, pages 63–70. Elsevier, 2015.

Multi-scale characterization of symbiont diversity in the pea aphid complex through metagenomic approaches

Authors : Cervin Guyomar[1,2] , Fabrice Legeai[1,2], Christophe Mougel[1], Claire Lemaitre[2], Jean-Christophe Simon[1]

1 : INRA, UMR 1349 INRA/Agrocampus Ouest/Université Rennes 1, Institut de Génétique, Environnement et Protection des Plantes (IGEPP), Le Rheu, France
2 : INRIA/IRISA/GenScale, Campus de Beaulieu, Rennes, France

Most metazoans are involved in durable symbiotic relationships with microbes which can take several forms, from mutualism to parasitism. The advances of NGS technologies and bioinformatics tools have opened new opportunities to shed light on this hidden but very influential diversity.
The pea aphid is a model insect system for symbiont studies. It harbors both an obligatory symbiont supplying key nutrients and several facultative symbionts bringing some novel functions to the host, such as protection against natural enemies and thermal stress. The pea aphid is organized in a complex of biotypes, each adapted to a specific host plant of the legume family and having its own symbiont composition. Yet, the metagenomic diversity of the biotype-associated symbiotic community is still largely unknown. In particular, little is known on how the symbiotic genomic diversity is structured at different scales: across host biotypes, amongst individuals of the same biotype, or within individual aphids.

We used high throughput whole genome metagenomic sequencing to characterize with a fine resolution the metagenomic diversity of both individual resequenced aphids and biotype specific pooled aphids. By a reference genome mapping approach, we first assessed the taxonomic diversity of the samples and built symbiont specific read sets. We then performed a genome-wide SNP-calling, to examine the differences in bacterial strains between samples. Our results revealed different diversity patterns at the three considered scales for the pea aphid symbionts. At the inter-biotype and intra-biotype scales, the primary symbiont *Buchnera* and some secondary symbionts such as *Serratia* showed a biotype specific diversity. We showed evidence for horizontal transfer of a *Hamiltonella* strain between biotypes, and found two distinct strains of *Regiella* symbionts within some biotypes. At the finest intra-host diversity scale, we also showed that these two strains of *Regiella* may coexist inside the same aphid host. This study highlights the huge potential of bioinformatics analyses of metagenomic dataset in exploring microbiote diversity in relation with host variation.
Keywords: metagenomics, aphids, diversity, symbionts

iMOMi: a database dedicated to integration and exploration of multi-omic data.

Nicolas PONS[1], Jean-Michel BATTO[2], Amine GHOZLANE[3], Kevin WEISZER[1], Pierre LEONARD[1], Emmanuelle LE CHATELIER[1], Pierre RENAULT[2] and S. Dusko EHRLICH[1]

[1] METAGENOPOLIS, INRA Université Paris-Saclay, 78350, Jouy-en-Josas, France
[2] MICALIS, INRA Université Paris-Saclay, 78350, Jouy-en-Josas, France
[3] Bioinformatics and Biostatistics Hub, Institut Pasteur, 75015, Paris France

Corresponding Author: nicolas.pons@inra.fr

With the explosion of data produced especially in the human metagenomic domain, numerous heterogeneous sources of data are available to the scientific community: metagenomic/metatranscriptomic reads, metagenome assemblies, reference gene catalogues, gene abundance profiles, sample metadata etc. In order to explore fully various facets of human microbiota, biologists need to access and browse integrated databases connecting these data. In this context, we developed an extension for iMOMi (integrative MultiOmic Mining relational database) in order to unify and centralize metagenomic reference gene catalogues with their accompanying data and provide a common datasource to biologists.

The iMOMi design is a modular database related to annotations data (Genbank, EMBL, COG/eggNOG, KEGG…) and *in silico* analysis (such as orthologous classification or regulatory motifs delineation) [1]. Current version of iMOMi is a relational database centered on the integration and exploration of (meta)genomic data. The database contains more than 150 tables with specific rules: annotation, phylogeny, taxonomy, metabolism, gene expression regulation, protein structure. iMOMi is capable to integrate genome data but also metagenomic data in particular large reference gene catalogues used for building abundance gene profiles in quantitative metagenomic studies. Furthermore, we added a powerful extension to handle (i) the concept of MetaGenomic Species; (ii) the relation between reference genes and their metagenome source and (iii) the gateways between several reference gene catalogues.

We developed user-friendly tools, for example iMOMi Studio which (i) facilitates the importation of new data (direct download of genomes from the NCBI, functional annotation of metagenome…) and (ii) allows a certain data browsing, in particular for metagenomic data: functional potential projection onto iPath2 [2] visualization for a genome/metagenome/MGS or taxonomic distribution with KRONA representation [3] for examples.

Furthermore, to help querying multiple tables, we developed a dedicated API (Application Programming Interface) based on PL/SQL stored procedures that can be called into any programming language as long as they can import SQL interface (Delphi, Python, Ruby, R…). The API contains functions for data integration, MGS content extraction, taxonomic distribution, functional extraction…

Finally, with the multiplication of reference gene catalogues increasingly reconstructed specifically for a quantitative metagenomic study, it becomes complicated for an investigator to cross the results of different studies. In this context, the implementation of iMOMi seeks to address this issue (i) by integrating numerous catalogues, (ii) by unifying their annotations (functional and taxonomic) and (iii) above all by bridging them. iMOMi have been optimized and is currently capable to handle dozen of catalogues representing hundreds millions of features.

The client software iMOMi Studio is available at http://mgps.eu/index.php?id=ibs-tools

References

[1] Pons N, Batto JM, Ehrlich SD, Renault P. Development of software facilities to characterize regulatory binding motifs and application to Streptococcaceae. *J Mol Microbiol Biotechnol.*14(1-3):67-73, 2008.
[2] Yamada T1, Letunic I, Okuda S, Kanehisa M, Bork P. iPath2.0: interactive pathway explorer. *Nucleic Acids Res.* Jul;39(Web Server issue):W412-5. 2011.
[3] Ondov BD1, Bergman NH, Phillippy AM. Interactive metagenomic visualization in a Web browser**.** *BMC Bioinformatics*. 12:385. 2011.

Sequence composition-based read binning and taxonomic profiling in infectious metagenomics diversity analyses

Mathias VANDENBOGAERT, Charlotte BALIÈRE and Valérie CARO

Pole for Genotyping of Pathogens, Laboratory for Urgent Response to Biological Threats, Environment and Infectious Risks Research and Expertise Unit, Institut Pasteur, 25-28 rue du docteur Roux, F-75724 Paris cedex 15, France

Corresponding Author: mathias.vandenbogaert@pasteur.fr

1 Introduction/Context

High-Throughput Sequencing (HTS) has gained in throughput and cost-efficiency, strongly affecting public health and biomedical research and enable to conduct large scale genomic projects. In infectious metagenomics studies, it is critical to detect rapidly and sensibly potential life-threatening pathogens. Our objective is to characterize the bacterial and viral composition inherent to clinical samples. Optimized methodologies for fast and accurate recovery of the known microbial biodiversity, followed with phylogenetic approaches to study the remaining ("dark matter", "BLAST-resistant") species are required to drive thorough sample comparisons in order to pinpoint both common and sample-specific infectious agents.

2 Discerning discriminative reads in complex metagenomics datasets

Clinical samples often contain either a number of (meta-) genomes that are at best divergently related to known references, or a limited number of genomes with very low coverage. Meta-genomic samples represent a large number of reads, rendering assembly without pre-condition computationally inefficient, and often proves to result in under-assembly and chimeric contigs. Variable levels of similarity challenges identification in diagnostics settings, where the distinction between presence and absence of single species or relative abundance levels are of eminent importance [1]. Sequence alignment through mapping proves difficult in such cases, and "genome finishing" often turns out to be impossible. To overcome the limitations of alignment-based comparisons, ultrafast alignment-free methods are exploited [2-4], providing a powerful comparison strategy to distinguish different organisms present in meta-genomics HTS read datasets.

We emphasize on clustering of sequencing reads, based on k-mer counts, as a preliminary and valuable step prior to assembly, as reads belonging to possibly different species show different k-mer compositions. Results indicate that, as alignment-free methods relying on clustering of word enumerations are obviously less accurate than direct sequence alignments, they should only be used when direct alignment is either impossible (due the high level of meta-genomic divergence) or computationally too complex. In addition, to be able to more thoroughly distinguish pathogen from host, the size of k has to be optimized as a function of the length of the reads. The length of the reads is indeed steadily increasing, with the different sequencing technologies at hand, i.e. the Illumina HiSeq/MiSeq technology generates 100 to 250 bp reads, whereas Life Technology's Ion Torrent PGM/Proton systems generate reads up to 300, 400 bp and above.

3 Taxonomic profiling and comparison of infectious metagenomics samples

As an extension, comparison of samples has the potential to recover the shared background diversity, and can shed light onto the microbiotic identity of samples, easing the pathogen identification task. K-mer frequencies are used either for taxonomic binning of individual reads or for computing the overall composition. K-mer distributions of a set of metagenomic samples give an indication of the presence, abundance and evolutionary relatedness of novel organisms present in the samples.

References

[1] Lindner MS and Renard BY. Metagenomic abundance estimation and diagnostic testing on species level. Nucl. Acids Res. (7 January 2013) 41 (1): e10.
[2] Vinga S and Almeida J: Alignment-free sequence comparison-a review. Bioinformatics 2003, 19(4):513–523.
[3] Song K, Ren J, Zhai Z, Liu X, Deng M and Sun F. Alignment-free sequence comparison based on next-generation sequencing reads. J Comput Biol 2013, 20(2):64–79.
[4] Trifonov V and Rabadan R. Frequency analysis techniques for identification of viral genetic data. MBio 2010, 1(3).

REGOVAR, logiciel libre pour l'analyse de données de séquençage haut débit pour les maladies génétiques rares

Anne-Sophie DENOMMÉ-PICHON[1], Olivier GUEUDELOT[2], Jérémie ROQUET[3], June SALLOU[4], Sacha SCHUTZ[5], David GOUDENÈGE[2]

[1] Laboratoire de génétique - UMR INSERM 954, CHRU de Nancy, 54500 Vandœuvre-lès-Nancy, France
[2] Laboratoire de génétique - UMR CNRS 6214-INSERM 1083, CHU d'Angers, 49933 Angers, France
[3] dnalyze.me, 75013 Paris, France
[4] Univ. Bordeaux, Centre de BioInformatique de Bordeaux (CBiB), 33000 Bordeaux, France
[5] Laboratoire de génétique, CHRU de Brest, 29609 Brest, France

Corresponding Author: asdp@regovar.org

1 Présentation

REGOVAR est un projet collaboratif, libre et ouvert de logiciel d'analyse de données de séquençage haut débit avec une interface graphique simple et conviviale pour les panels de gènes, l'exome et le génome (DPNI, recherche de SNV, CNV, SV…). Le projet est financé dans le cadre d'un PHRCI du Grand Ouest (Angers, Brest, Nantes, Poitiers, Rennes et Tours) pour structurer les généticiens cliniciens, biologistes et bioinformaticiens impliqués dans le diagnostic moléculaire des maladies génétiques rares. Si la bioinformatique médicale appliquée au NGS permet aujourd'hui d'analyser avec succès un grand nombre de données au sein d'un CHU ou d'une région, elle souffre d'un manque de coordination à l'échelle nationale. REGOVAR vise à impliquer et fédérer les différentes communautés concernées, sans limites institutionnelles ou géographiques. Il se base exclusivement sur des technologies et des logiciels libres et gratuits, éliminant toute contrainte contractuelle et budgétaire.

2 Objectifs

REGOVAR est un logiciel permettant le traitement de données génétiques, depuis la récupération des fichiers produits par les séquenceurs, quelle qu'en soit la technologie, jusqu'à la génération de rapports illustrés et de comptes-rendus d'analyse en passant par les contrôles de qualité, la détection, l'annotation, le filtrage, la priorisation et la visualisation de variants. Son architecture client-serveur permet une utilisation depuis des ordinateurs de bureau sous Windows, Linux et macOS, via une interface graphique claire et intuitive conçue pour permettre l'analyse des données par des généticiens n'ayant pas de compétence spécifique en bioinformatique (filtres de variants enregistrables, simplification de la bioanalyse…). Sa conception modulaire permet d'intégrer dynamiquement de nouveaux pipelines, quelles que soient leurs dépendances, qui peuvent être partagés au sein de la communauté. Ces échanges de pipelines permettront à terme l'harmonisation des bonnes pratiques avec des pipelines unifiés nationalement et validés par l'ANPGM. REGOVAR intègre une base de données principale dimensionnée pour supporter aussi bien l'analyse de panels, d'exomes, que de génomes complets. Cette base est enrichie de données publiques telles que celles provenant de gnomAD et dbNSFP, ainsi que de données locales. Des échanges anonymisés de certaines informations recueillies sont possibles. REGOVAR permet une utilisation aussi bien en recherche qu'en diagnostic.

3 Appel à collaboration

REGOVAR a déjà suscité l'intérêt d'autres acteurs comme les laboratoires de génétique des CHU de Nancy et de Montpellier. Le projet est ouvert à toute personne souhaitant apporter sa contribution : idées, intégration de pipeline, développement, test, documentation… Informations disponibles sur https://regovar.org/.

Abréviations

ANPGM : Association Nationale des Praticiens de Génétique Moléculaire

DPNI : Dépistage Prénatal Non Invasif

PHRCI : Projet Hospitalier de Recherche Clinique Interrégional

Alignement à grande échelle pour une approche métagénomique dans le cadre du projet Tara *Oceans*

Artem KOURLAIEV[1], Corinne DA SILVA[1], Stefan ENGELEN[1], Alexis BERTRAND[1], Aimeric BRUNO[1], Eric PELLETIER[2] Patrick WINCKER[2] and Jean-Marc AURY[1]

[1] Commissariat à l'Energie Atomique (CEA), Institut François Jacob, Genoscope, 91000 Evry, France.
[2] CEA / Genoscope, CNRS UMR 8030, Université d'Evry, France.

Corresponding Author: artem.kourlaiev@cea.fr

Le projet *Tara Oceans* (2009-2013) a pour objectif d'étudier globalement les écosystèmes planctoniques marins (des virus aux métazoaires) [1]. Des échantillons d'eau ont été prélevés puis filtrés sur 88 stations couvrant l'ensemble des océans [2]. Le séquençage de l'ADN et de l'ARN présents dans ces échantillons a été effectué au Genoscope. Ce projet constitue le plus grand effort de séquençage jamais réalisé pour des organismes marins.

La métagénomique est une méthode permettant d'étudier l'ensemble des génomes (et des transcriptomes pour la métatrancriptomique) des populations de micro-organismes d'un écosystème donné à partir d'un échantillon environnemental. C'est l'une des approches choisies pour l'étude de la biodiversité des océans dans le cadre du projet *Tara Oceans*.

Pour l'étude des eucaryotes présents dans ces échantillons, un catalogue de gènes a été constitué à partir des données métatranscriptomiques. Plus de 116 millions de gènes ont été identifiés. Afin de déterminer les niveaux d'abondance et d'expression de chacun de ces gènes à chaque station de prélèvement, il s'est révélé nécessaire d'aligner l'ensemble des lectures métagénomiques et métatranscriptomiques sur cette référence. Ceci apporte un premier niveau de valorisation aux données et offre des nouvelles voies d'analyses.

La volumétrie du catalogue de gènes et de lectures à aligner (plus de 460 milliards) a nécessité la mise en place d'une méthode d'alignement massive, en plusieurs étapes : dans un premier temps, la référence a été fragmentée et l'ensemble des échantillons ont été alignés sur chacun des fragments. Ensuite, des références de plus petite taille (contenants uniquement les gènes avec au moins un match obtenu lors du premier alignement) ont été constituées pour chaque échantillon. Enfin, un second alignement du même ensemble d'échantillons a été effectué sur ces références réduites.

Grâce à une conception massivement parallèle et à l'utilisation des moyens de calculs du TGCC-CCRT [3], des dizaines de milliers de cœurs ont pu être utilisés en parallèle pour diminuer le temps de restitution. Nous avons ainsi pu obtenir l'ensemble des résultats en une semaine. La méthode mise en place est extensible, et pourra être utilisée pour d'autres projets de métagénomique.

Références

[1] Karsenti, E., S. G. Acinas, et al. A holistic approach to marine eco-systems biology. *PLoS Biol* 9(10): e1001177, 2011.
[2] Pesant, S. *et al.* Open science resources for the discovery and analysis of Tara *Oceans* data. *Sci. Data* **2**, 150023 (2015).
[3] Très Grand Centre de Calcul du CEA, http://www-hpc.cea.fr/fr/complexe/tgcc.htm

Seqenv: linking sequences to environments through text mining

Lucas Sinclair[1,†], Umer Z Ijaz[2,†], Lars Juhl Jensen[3], Marco Coolen[4], Cecile Gubry-Rangin[5], Alica Chrokov[6], Anastasis Oulas[7,8], Christina Pavloudi[7], Julia Schnetzer[9], Aaron Weimann[10], Ali Zeeshan Ijaz[11], Alexander Eiler[1], Christopher Quince[12,*], Evangelos Pafilis[7,*].

1 Department of Ecology and Genetics, Limnology, Uppsala University, Sweden
2 Infrastructure and Environment Research Division, School of Engineering, University of Glasgow, United-Kingdom
3 The Novo Nordisk Foundation Center for Protein Research, Faculty of Health and Medical Sciences, University of Copenhagen, Denmark
4 Western Australia Organic and Isotope Geochemistry Centre (WA-OIGC), Department of Chemistry, Bentley Campus, Curtin University, Australia
5 Institute of Biological & Environmental Sciences, University of Aberdeen, Scotland, United Kingdom
6 Biology Centre of the Czech Academy of Sciences, Institute of Soil Biology, Czech Republic
7 Institute of Marine Biology Biotechnology and Aquaculture (IMBBC), Hellenic Centre for Marine Research (HCMR), Heraklion Crete, Greece
8 Bioinformatics Group, The Cyprus Institute of Neurology and Genetics, Nicosia, Cyprus **9** Max Planck Institute for Marine Microbiology, Department of Molecular Ecology, Microbial Genomics and Bioinformatics Group, Bremen, Germany
10 Computational Biology of Infection Research, Helmholtz Centre for Infection Research, Braunschweig, Germany
11 Hawkesbury Institute for the Environment, Western Sydney University, Hawkesbury, Australia
12 Warwick Medical School, University of Warwick, Coventry, United-Kingdom
† These authors contributed jointly to this work.
* Corresponding authors: C. Quince, c.quince@warwick.ac.uk
* Address: Warwick Medical School, University of Warwick, Coventry CV4 7AL, U.K.
* Corresponding authors: E. Pafilis, pafilis@hcmr.gr
* Address: Institute of Marine Biology Biotechnology and Aquaculture, Hellenic Centre for Marine Research, P.O. Box 2214, Heraklion, 71003, Crete, Greece.

Corresponding author: c.quince@warwick.ac.uk

1 Abstract

Understanding the distribution of taxa and associated traits across different environments is one of the central questions in microbial ecology. High-throughput sequencing (HTS) studies are presently generating huge volumes of data to address this biogeographical topic. However, these studies are often focused on specific environment types or processes, leading to the production of individual unconnected datasets. The large amounts of legacy sequence data with associated metadata that exist can be harnessed to better place the genetic information found in these surveys into a wider environmental context. Here we introduce a software program, `seqenv`, to carry out precisely such a task. It automatically performs similarity searches of short sequences against the "nt" nucleotide database provided by NCBI and, out of every hit, extracts – if it is available – the ¡isolation source¿ textual metadata field. After collecting all the isolation sources from all the search results, we run a text mining algorithm to identify and parse words that are associated with the Environmental Ontology (EnvO) controlled vocabulary. This, in turn, enables us to determine both in which environments individual sequences or taxa have previously been observed and, by weighted summation of those results, to summarize complete samples. We present two demonstrative applications of `seqenv` to a survey of ammonia oxidizing archaea as well as to a plankton paleome dataset from the Black Sea. These demonstrate the ability of the tool to reveal novel patterns in HTS and its utility in the fields of environmental source tracking, paleontology, and studies of microbial biogeography. To install `seqenv`, go to: https://github.com/xapple/seqenv.

Predicting the ecological quality status of marine environments from eDNA metabarcoding data using supervised machine learning

Tristan CORDIER[1], Philippe ESLING[2], Franck LEJZEROWICZ[1], Joana VISCO[3], Amine OUADAHI[1], Catarina MARTINS[4], Tomas CEDHAGEN[5] and Jan PAWLOWSKI[1,3]

[1] Department of Genetics and Evolution, University of Geneva, Boulevard d'Yvoy 4, CH 1205 Geneva, Switzerland

[2] IRCAM, UMR 9912, Université Pierre et Marie Curie, 4 place Jussieu, 75005 Paris, France

[3] ID-Gene ecodiagnostics, Ltd, chemin des Aulx 14, 1228 Plan-les-Ouates, Switzerland

[4] Marine Harvest ASA, Sandviksboder 77AB, Bergen, 5035 Bergen, Norway

[5] Department of Bioscience, Section of Aquatic Biology, University of Aarhus, Nordre Ringgade 1, 8000 Aarhus, Denmark

Corresponding Author: tristan.cordier@unige.ch

Abstract

Monitoring biodiversity is essential to assess the impacts of increasing anthropogenic activities in marine environment. Traditionally, marine biomonitoring involves the sorting and morphological identification of benthic macro-invertebrates, which is time-consuming and taxonomic-expertise demanding. High-throughput amplicon sequencing of environmental DNA represents a promising alternative for benthic monitoring. However, an important fraction of eDNA sequences remains unassigned or belong to taxa of unknown ecology, which prevent their use for assessing the ecological quality status. Here, we show that supervised machine learning (SML) can be used to build robust predictive models for benthic monitoring, regardless the taxonomic assignment of eDNA sequences. We tested three SML approaches to assess the environmental impact of marine aquaculture using benthic foraminifera eDNA, a group of unicellular eukaryotes known to be good bioindicators. We found similar ecological status as obtained from morphology-based macrofaunal inventories. We argue that SML approaches could overcome and even bypass the cost and time-demanding morpho-taxonomic approaches in future biomonitoring.

Extended HLA haplotypes in Membranous Nephropathy

LOUIS ÉDOUARD CHAUVIÈRE[1], PIERRE RONCO[1] and HANNA DEBIEC[1]

[1] UMR-S 1155, Hôpital Tenon, 4 rue de la Chine, Bât. Recherche, 75020 PARIS

Corresponding Author: louis.chauviere@inserm.fr

Membranous Nephropathy (MN) is an Auto-immune and multifactorial disease caused by the production of PLA2R1 antigens in podocytes [1] (specialized cells for filtration in kidney), resulting in the activation of the immune response and podocyte destruction. Recently, using genomics studies, risk variants were found on PLA2R1 gene [2] and HLA-II cluster, namely on genes coding for the antigen implicated in MN and for proteins implicated in antigen presentation.

Data analysis was focused on the three HLA regions using Whole Exome Sequencing data from 123 patients. The HLA-II risk haplotype in Caucasian European population [3] was first confirmed as HLA-DRB1*03:01, HLA-DQA1*05:01, HLA-DQB1*02:01 (known as DR3-DQ2 haplotype). This haplotype minor allele frequency is 0.29 in the MN cohort versus 0.09 in the control Caucasian European population.

The aim was to fully explore the entire HLA cluster and to find new MN risk factors. To search for epistatic variant interactions, we have used regression-based methods. This analysis demonstrated that the haplotype can be extended with HLA-I allelotypes : HLA-A*01:01, HLA-B*08:01, HLA-C*07:01 (A1-B8-DR3-DQ2). Between those two cluster, HLA-III groups numerous different types of genes that are important in the Immune Response as C2, C4A and C4B (complement components), CFB (Complement Factor B) or TNF.

In HLA-III region, the analysis was focused on paralogous C4A and C4B genes. Only five nucleotides on exon 26 are different between the two genes. Each one can possess an endogenous retrovirus (HERV-K(C4)) [4], located in intron 9, resulting in a longer protein (C4L). In addition, Copy Number Variations (CNVs) were found on each gene. CNVs were detected aligning C4A and C4B reads on C4A reference sequence. Then, the mean depth of coverage on exon 26 was calculated to establish C4 copy number. C4A and C4B were split using the proportion of variants that identify the two genes. An interaction was detected between the loss of C4A and the extended haplotype A1-B8-DR3-DQ2. Variants located in HLA-I genes as DDR1, VARS2, SFTA, TRIM26 or HCG17, interact with the loss of C4A and could be added to this extended haplotype.

The risk factor for MN is finally not only linked with DR3-DQ2 haplotype in HLA-II region, but with an extended haplotype covering all of the three HLA regions. The next step will be to explore intronic and intergenic regions, especially HERV-K(C4).

References

[1] Laurence H. Beck, Ramon G.B. Bonegio, Gérard Lambeau, David M. Beck, David W. Powell, Timothy D. Cummins, Jon B. Klein, David J. Salant. M-type phospholipase A2 receptor as target antigen in idiopathic membranous nephropathy. *N Engl J Med*, 361: 11–21, 2009

[2] Horia C. Stanescu, Mauricio Arcos-Burgos, Alan Medlar, Detlef Bockenhauer, Anna Kottgen, Liviu Dragomirescu, Catalin Voinescu, Naina Patel, Kerra Pearce, Mike Hubank, Henry A.F. Stephens, Valerie Laundy, Sandosh Padmanabhan, Anna Zawadzka, Julia M. Hofstra, Marieke J.H. Coenen, Martin den Heijer, Lambertus A.L.M. Kiemeney, Delphine Bacq-Daian, Benedicte Stengel, Stephen H. Powis, Paul Brenchley, John Feehally, Andrew J. Rees, Hanna Debiec, Jack F.M. Wetzels, Pierre Ronco, Peter W. Mathieson, Robert Kleta. Risk HLADQA1 and PLA(2)R1 alleles in idiopathic membranous nephropathy. *N Engl J Med*, 364:616–626, 2011

[3] Peggy Skula, Yong Li, Horia C. Stanescu, Matthias Wuttke, Arif B. Ekici, Detlef Bockenhauer, Gerd Walz, Stephen H. Powis, Jan T. Kielstein, Paul Brenchley, GCKD Investigators, Kai-Uwe Eckardt, Florian Kronenberg, Robert Kleta, Anna Köttgen. Genetic risk variants for membranous nephropathy: extension of and association with other chronic kidney disease aetiologies. *Nephrol Dial Transplant*, 32 (2): 325-332, 2017.

[4] Mike J. Mason, Cate Speake, Vivian H. Gersuk, Quynh-Anh Nguyen, Kimberly K. O'Brien, Jared M. Odegard, Jane H. Buckner, Carla J. Greenbaum, Damien Chaussabel and Gerald T. Nepom. Low HERV-K(C4) copy number is associated with Type 1 Diabetes. *Diabetes*, 63(5): 1789-1795, 2014.

RSAT matrix-clustering: dynamic exploration and redundancy reduction of transcription factor binding motif collections

Jaime Abraham CASTRO-MONDRAGON[1], Sébastien JAEGER[2], Denis THIEFFRY[3], Morgane THOMAS-CHOLLIER[3,*] and Jacques VAN HELDEN[1,*]

[1] Aix-Marseille Univ, Inserm, TAGC, Technological Advances for Genomics and Clinics, UMR_S 1090, Marseille, France.

[2] Aix Marseille Univ, CNRS, INSERM, CIML, Marseille, France

[3] IBENS, Département de Biologie, Ecole Normale Supérieure, CNRS, Inserm, PSL Research University, F-75005 Paris, France

* Corresponding Authors: Jacques.van-Helden@univ-amu.fr , mthomas@biologie.ens.fr

Transcription Factor (TF) databases contain multitudes of binding motifs (TFBMs) from various sources, from which non-redundant collections are derived by manual curation. The advent of high-throughput methods stimulated the production of novel collections with increasing numbers of motifs. Meta-databases, built by merging these collections, contain redundant versions, because available tools are not suited to automatically identify and explore biologically relevant clusters among thousands of motifs. Motif discovery from genome-scale data sets (e.g., ChIP-seq) also produces redundant motifs, hampering the interpretation of results. We present *matrix-clustering* [1], a versatile tool that clusters similar TFBMs into multiple trees, and automatically creates non-redundant TFBM collections. A feature unique to *matrix-clustering* is its dynamic visualisation of aligned TFBMs, and its capability to simultaneously treat multiple collections from various sources. We demonstrate that *matrix-clustering* considerably simplifies the interpretation of combined results from multiple motif discovery tools, and highlights biologically relevant variations of similar motifs. We also ran a large-scale application to cluster ~11,000 motifs from 24 entire databases, showing that *matrix-clustering* correctly groups motifs belonging to the same TF families, and drastically reduced motif redundancy. *matrix-clustering* is integrated within the RSAT suite (http://rsat.eu/) [2], accessible through a user-friendly web interface or command-line for its integration in pipelines.

Availability: http://rsat.eu/

Manuscript in press in Nucleic Acids Research.

Preprint accessible on bioRxiv: doi: https://doi.org/10.1101/065565

References

[1] Jaime A. Castro-Mondragon, Sébastien Jaeger, Denis Thieffry, Morgane Thomas-Chollier# and Jacques van Helden#, RSAT matrix-clustering: dynamic exploration and redundancy reduction of transcription factor binding motif collections. *Nucleic Acids Research*, in press.

[2] Medina-Rivera A*, Defrance M*, Sand O*, Herrmann C, Castro-Mondragon J, Delerce J, Jaeger S, Blanchet C, Vincens P, Caron C, Staines DM, Contreras-Moreira B, Artufel M, Charbonnier – Khamvongsa L, Hernandez C, Thieffry D, Thomas-Chollier M#, van Helden J# (2015) RSAT 2015: Regulatory Sequence Analysis Tools. *Nucleic Acid Research* **43**: W50-6.

Assemblage de novo de quasi-espèces virales basé sur un graphe de chevauchements

Jasmijn A. Baaijens[1], Amal Zine El Aabidine[2,3], Eric Rivals[2,3,†] et Alexander Schönhuth[†1,†]
[1] Centrum Wiskunde & Informatica, Amsterdam, Netherlands
[2] LIRMM, CNRS and Université de Montpellier, Montpellier, France
[3] Institut Biologie Computationnelle, CNRS and Université de Montpellier, France

Auteur référent: rivals@lirmm.fr

[†] Ces auteurs ont des contributions égales et sont tous deux auteurs correspondants.

Durant une épidémie, les virus qui infectent des humains (ou d'autres espèces) doivent échapper au système immunitaire de l'hôte. Pour cela, leur génome mute rapidement et cela engendre une diversité des virus dont le comportement évolue – on parle alors de quasi-espèces virales. Grâce à ces mutations génomiques, certains de ces virus pourront mieux combattre le système immunitaire ou mieux résister à un traitement. On ne sait pas à l'avance quelles quasi-espèces s'adaptent le mieux à l'hôte, ni lesquelles deviennent majoritaires dans la population de virus d'un individu ; la fréquence relative des quasi-espèces est une inconnue importante du point de vue biologique et médical. Pour lutter contre les épidémies ou suivre les effets d'un traitement, on doit savoir quelles quasi-espèces acquièrent des mutations avantageuses, et il est donc crucial d'obtenir les séquences de leurs génomes et d'estimer leurs proportions dans la population virale d'un individu. On peut séquencer les génomes viraux présents chez l'hôte, grâce aux techniques de séquençage à haut débit qui produisent des millions de lectures courtes (*reads* en anglais). Il faut ensuite assembler ces fragments pour obtenir les séquences génomiques des quasi-espèces. Bien que de nombreux outils bioinformatiques d'assemblage de génome existent, ils sont inadaptés au cas des quasi-espèces pour lequel le logiciel doit deviner combien de quasi-espèces sont présentes, puis reconstruire partiellement ou entièrement le génome de chacune, afin d'identifier leurs différences. Les méthodes existantes reposent sur une séquence de référence du génome de l'espèce virale. Malheureusement, on dispose rarement d'un génome de référence, en particulier lors d'épidémie ou lors de l'apparition de nouveaux virus (par exemple dans le cas de zoonoses). Il faut donc considérer le cas dit *de novo* où aucune séquence de référence n'est disponible. Une des difficultés est d'assembler des génomes similaires tout en les distinguant (en séparant les quasi-espèces). Nous proposons une méthode nommée SAVAGE pour assembler les génomes de quasi-espèces virales lorsque l'on ne dispose pas déjà d'un génome de référence (le cas le plus difficile d'assemblage). SAVAGE se base sur un graphe de chevauchement (ou *overlap graph* en anglais) qu'il calcule grâce à un algorithme performant. Les tests sur des données de HIV, des virus de l'hépatite C, des virus de Zika et d'Ebola démontrent la capacité de SAVAGE à reconstruire les quasi-espèces et à estimer leur fréquence relative [1]. Ce travail offre des perspectives nouvelles pour le suivi des infections chez les patients par des approches basées sur le séquençage à haut débit.

SAVAGE est accessible à https://bitbucket.org/jbaaijens/savage.

Remerciements

Les travaux de AS sont financés par Vidi grant 639.072.309, de the Netherlands Organisation for Scientific Research (NWO). Ceux de ER, AZ sont financés par l'Institut de Biologie Computationnelle (ANR-11-BINF-0002), France Génomique, et le Défi Mastodons C3G.

Références

[1] Jasmijn A. Baaijens, Amal Zine El Aabidine, Eric Rivals, and Alexander Schönhuth. De novo assembly of viral quasispecies using overlap graphs. *Genome Research*, 27(5) :835–848, 2017.

Caractérisation de données biologiques

Arthur CHAMBON[1], Tristan BOUREAU[2], Frédéric LARDEUX[1]and Frédéric SAUBION[1]
[1] LERIA, Université d'Angers, France
[2] IRHS, Université d'Angers, France

Corresponding author: arthur.chambon@univ-angers.fr

L'analyse logique de données [1,2,3] constitue une alternative originale aux approches traditionnelles de classification de données issues de l'apprentissage artificiel. L'objectif principal de cette méthodologie consiste à se concentrer sur la justification explicite de la classification de données dans des groupes/classes. Considérons un ensemble Ω d'observations, dont chaque élément est représenté par des données appartenant à un ensemble d'attributs Booléens $\mathcal{A}$. Cet ensemble d'observations étant divisé en deux groupes (groupe positif P et groupe négatif N), l'analyse logique de données (LAD) consiste à trouver des "patterns" (motifs) caractérisant un groupe, i.e., des expressions booléennes sur un sous-ensemble de $\mathcal{A}$ vérifiées par au moins une observation d'un groupe, et chez aucune observation de l'autre groupe.

Le problème de caractérisation multiple de données (MCP) [4,5] consiste également à caractériser les observations. Toutefois, il diffère de l'approche LAD sur différents points:

— Une solution du MCP sera non pas un pattern, mais un sous ensemble d'attributs.
— Une instance du MCP peut contenir plus de deux groupes. Une solution au MCP caractérisera tous les groupes simultanément.

Il s'agit donc de déterminer un ensemble d'attributs, caractérisant chacun des groupes.
Nous pouvons ainsi représenter ces données par une matrice telle que :

— Chaque ligne i représente les observations.
— Chaque colonne j représente les attributs.
— La valeur x_{ij} vaut 1 si l'attribut j est présent chez l'observation i, 0 sinon.

Observations	Attributs / Groupes	a	b	c	d	e
1	1	0	1	1	1	1
2	1	1	0	1	1	1
3	1	0	1	1	1	1
4	2	0	0	1	1	0
5	2	1	1	1	0	1
6	3	1	1	0	1	1

Nous déterminons l'ensemble des solutions optimisant deux critères :

— Minimiser le nombre d'attributs dans les solutions afin de déterminer le groupe auquel appartiennent les observations à moindre coût.
— Maximiser la similarité des observations de même groupe afin de mieux analyser les différents groupes.

References

[1] Gabriela Alexe, Sorin Alexe, Tibérius O. Bonates, and Alexander Kogan. Logical analysis of data - the vision of peter l. hammer. *Ann. Math. Artif. Intell.*, 49(1-4):265–312, 2007.

[2] Endre Boros, Yves Crama, Peter L. Hammer, Toshihide Ibaraki, Alexander Kogan, and Kazuhisa Makino. Logical analysis of data: classification with justification. *Annals OR*, 188(1):33–61, 2011.

[3] Peter L. Hammer and Tibérius O. Bonates. Logical analysis of data - an overview: From combinatorial optimization to medical applications. *Annals OR*, 148(1):203–225, 2006.

[4] Fabien Chhel, Frédéric Lardeux, Adrien Goëffon, and Frédéric Saubion. Minimum multiple characterization of biological data using partially defined boolean formulas. In *Proceedings of the ACM Symposium on Applied Computing, SAC 2012, Riva, Trento, Italy, March 26-30, 2012*, pages 1399–1405, 2012.

[5] Arthur Chambon, Tristan Boureau, Frédéric Lardeux, Frédéric Saubion, and Marion Le Saux. Characterization of multiple groups of data. In *Tools with Artificial Intelligence (ICTAI), 2015 IEEE 27th International Conference on*, pages 1021–1028. IEEE, 2015.

Conservation of interaction energy landscapes across structural homologs through cross-docking calculations

Hugo SCHWEKE[1], Sophie SACQUIN-MORA[2], Marie-Hélène MUCCHIELLI-GIORGI[1,3] and Anne LOPES[1]

[1] Institute for Integrative Biology of the Cell (I2BC), CEA, CNRS, Univ. Paris-Sud, Université Paris-Saclay, 91198 Gif-sur-Yvette cedex, France
2 Laboratoire De Biochimie Théorique, CNRS UPR 9080, Institut De Biologie Physico-Chimique, Paris, France
[3] Sorbonne Universités, UPMC Univ Paris 06, UFR927, F-75005, Paris, France

Corresponding Author: anne.lopes@u-psud.fr

In 2003, Aloy *et al*. showed that homologs sharing at least 30% sequence identity almost invariably interact the same way [1]. In this study, we ask whether the whole interaction energy landscape of two interacting partners is conserved during evolution. Particularly, are the low-energy binding sites on a protein surface, as well as the less energetically favourable regions conserved among structural homologs? We tested this hypothesis through a large scale cross-docking experiment where structural homologs and arbitrary ligands were docked with a same receptor. We constituted a database of 72 protein structures divided into 12 structural homolog families. We performed a complete cross-docking experiment with ATTRACT [2]. Each protein played alternately the role of the receptor and of the ligand. For each pair of proteins, we produced a two dimensional (2D) energy map reflecting its docking energy landscape. We compared the energy maps of a receptor docked with all the ligands and ask whether the energy maps produced by structurally related ligands are more similar than those produced by unrelated ones.

This experiment highlights three major results: (i) for each receptor of the database, docking energy landscapes are more similar for ligands belonging to the same family. To quantify this effect, we measured our capacity to retrieve the corresponding families of the 72 ligands from the classification of their energy maps solely. The resulting Area Under the Curve (AUC) value is 0.83 and reflects clearly that structural homologs share similar interaction energy landscapes when interacting with a same receptor. (ii) the classification reveals that four structural families are subdivided into two sub-clusters that produce clearly distinct docking energy landscapes. This split cannot be explained by classical descriptors such as sequence identity or RMSD. Instead this is mainly explained by different distributions of charges at the surfaces of homologs. Interestingly, these subdivisions seem to reveal particular biophysical or functional properties of these proteins. (iii) to distinguish regions favourable to interactions from unfavourable ones, we discretized the 2D energy maps into distinct energy classes. This representation enables to pinpoint interaction hot-spots, warm and cold spots. Interestingly we show that the information provided by either the warm and cold spots is sufficient to correctly classify the ligands suggesting that not only the hot spots but also the rest of the surface have been constrained during evolution.

These results show that (i) the whole docking energy landscape seems to be conserved among structural homologs, (ii) protein docking can highlight biophysical and functional properties of structural homologs that could not be revealed by classical descriptors and (iii) warm and cold spots contain important information on the properties of a protein family. These regions may play an important role in protein interaction by competing with the effective native binding site.

References

[1] Aloy P, Ceulemans H, Stark A, Russell RB. The relationship between sequence and interaction divergence in proteins. J Mol Biol., 332:989-98, 2003.

[2] Fiorucci S, Zacharias M. Binding site prediction and improved scoring during flexible protein-protein docking with ATTRACT. Proteins, 78:3131-9, 2010.

Dynamic model of Central Carbon Metabolism and electrochemical reactions of the cell.

Cécile MOULIN[1,2], Jorgelindo DA VEIGA MOREIRA[3], Erwan BIGAN[3], Laurent SCHWARTZ[4], Mario JOLICOEUR[5], Laurent TOURNIER[2] and Sabine PERES[1,2]

[1] LRI, Universite Paris-Sud & UMR CNRS 8623, F-91405 Orsay, France
[2] MaIAGE, INRA, Universite Paris-Saclay, 78350 Jouy-en-Josas, France
[3] Laboratoire d'Informatique du l'X, UMR 7161, Ecole Polytechnique, F-91128 Palaiseau, France
[4] Assistance Publique des Hopitaux de Paris, F-75015 Paris, France
[5] Research Laboratory in Applied Metabolic Engineering, Department of Chemical Engineering, Ecole Polytechnique de Montreal, C.P. 6079, Centre-ville Station, Montreal (Quebec), Canada

Corresponding author: cecile.moulin@lri.fr

Abstract

In order to better understand cell proliferation, we propose to study several well-known intracellular oscillators such as NAD(H), pH, ATP, NADP(H). To understand and explain some experimental data [1] that have been obtained from fresh normal and cancer cells, extracted from human colon after a colectomy, we have contructed a dynamical model which synergistically combines two layers: a metabolic model [2] and an electrochemical model [3]. The metabolic part is focused on central carbon metabolism (CCM), since it represents the metabolic fingerprint of the cell that intimately interconnects with all cellular functions and intrinsic regulatory mechanisms. The electrochemical part represents the set of ionic reactions impacting the pH of the cell. In this work, the CCM was anchored to the electrochemical dynamics by means of the intracellular protons (H+) and energy (ATP-ADP) management.

References

[1] J Da Veiga Moreira. The Redox Status of Cancer Cells Supports Mechanisms behind the Warburg Effect. *Metabolites*, 6(4):1–12, 2016.

[2] Julien Robitaille, Jingkui Chen, and Mario Jolicoeur. A single dynamic metabolic model can describe mab producing cho cell batch and fed-batch cultures on different culture media. *PLOS ONE*, 10(9):1–23, 2015.

[3] Yann Bouret, Médéric Argentina, and Laurent Counillon. Capturing intracellular ph dynamics by coupling its molecular mechanisms within a fully tractable mathematical model. *PLOS ONE*, 9(1):1–11, 2014.

Proposition d'un workflow d'analyse QIIME dans Galaxy et évaluation de trois techniques d'extraction d'ADN pour l'analyse du microbiote intestinal 16S

S. BUFFET BATAILLON[1], M. BEN ABDALLAH[1], P. BORDRON[2], E. CORRE[2], S. KAYAL[1]

[1] Service Bactériologie, Hygiène hospitalière, Centre Hospitalier Universitaire de Rennes, 2 rue Henri Le Guilloux, 35033, Rennes, France

[2] CNRS - FR2424 - ABiMS, Station Biologique de Roscoff, Place Georges Teissier, 29680, Roscoff, France

Corresponding Author: marouane.bah@gmail.com

Introduction

Le microbiote intestinal est composé de l'ensemble des microorganismes présents dans le tractus digestif. Il est aujourd'hui identifiable grâce à une révolution biotechnologique : le séquençage à haut débit (NGS). Cette technique permet d'identifier jusqu'à 500 OTU (Operational Taxonomic Units) ou espèces bactériennes dans un microbiote intestinal. Elle nécessite différentes étapes que sont l'extraction, l'amplification, le séquençage d'ADN et le traitement des données. L'extraction d'ADN peut être faite sur colonne par des kits manuels comme Powerfecal Mobio® (Gold standard) (PWF), ou par des appareils automatiques sur billes magnétiques comme MagNapure 2.0 Roche® (MA) et Qiasymphony Qiagen®(QIA). L'amplification a lieu dans la zone V3-V4 de la région de l'ARN ribosomal 16S présente chez toutes les bactéries.

Objectifs et méthodes

Notre objectif principal est de créer et valider un workflow complet d'analyse des données de reads s'appuyant sur le package *Quantitative Insights Into Microbial Ecology* (QIIME) au sein d'une instance Galaxy. Notre objectif secondaire est d'analyser avec ce worflow et RStudio (RStudio Inc) l'impact de l'extraction d'ADN par PWF, MA et QIA sur les résultats NGS du microbiote 16S au CHU de Rennes.

Résultats

Le workflow QIIME sur Galaxy que nous proposons comporte 10 étapes principales de traitement de données allant de la préparation des séquences jusqu'au calcul de l'alpha et de la beta diversité de nos échantillons. Plus précisément, nous avons tout d'abord (1) décrit nos échantillons (*Validate mapping file)*. Nous avons ensuite (2) sélectionné des amplicons d'intérêt (*Split Fastq Libraries)*, puis nous avons (3) regrouper les séquences similaires en OTU (*Pick OTU)* avec un seuil de similarité de 0.97. (4) La séquence la plus abondante de chaque OTU obtenu a ensuite été choisi comme séquence représentative de l'OTU et les OTU ont été filtrés selon leur abondance (*Pick rep set)*. (5) nous avons ensuite aligné les séquences représentatives (*Align sequences*) selon les méthodes Pynast et Uclust pour ensuite (6) leur assigner une taxonomie issue de la base Greengenes (*Assign taxonomy)*. A partir de ce moment-là, (7) nous avons pu calculer la répartition du nombre d'OTU par échantillon (*Make OTU table)* tout en prenant soin de retirer les OTU non pertinents (*Filters OTU from OTU table)*. (8) Nous avons également présenté nos OTU selon la distance euclidienne (*Make phylogeny)* . (9) Nous avons ensuite estimé l'alpha diversité selon les indices de Chao1, Shannon et Simpson et (10) la beta diversité selon les distances Unifrac et Weighted Unifrac en utilisant respectivement les outils *Calculate Alpha diversity* et *Calculate Beta diversity*. L'ADN de 5 échantillons de selles a été extrait pour un même patient : 1 par PWF , 2 par MA , 2 par QIA. La table d'occurrence des OTU était identique quelque soit la technique d'extraction d'ADN utilisée. Aucune différence significative n'était montrée en analyse en composante principale entre les 3 méthodes d'extraction d'ADN (Indice KMO =0,72 et test de Bartlett <0,001). Les résultats d'alpha diversité et de bêta diversité étaient significativement comparables par tests ANOVA, par la méthode ADONIS.

Conclusion

Notre workflow est opérationnel sur la plateforme Galaxy ABIMS de la Station Biologique de Roscoff. Il nous a permis de montrer que notre méthode d'extraction automatique d'ADN MagNapure 2.0 Roche® peut être utilisée comme technique d'extraction d'ADN pour de l'analyse de microbiote 16S par NGS au CHU de Rennes.

BIOSPECIMENS v1.5: a web platform to facilitate collaborative research on infectious diseases

Karen LOUIS[1], Blandine RIMBAULT[2], Clément DELESTRE[1], Yoann MOUSCAZ[1], Marine ALBRIEUX[1], Christelle BOISSE[1], Régis VILLET[1], Guillaume BOISSY[1]

[1] BIOASTER, 40 avenue Tony Garnier, 69007, Lyon, France
[2] BIOASTER Paris, Bâtiment F. Jacob – 28 rue du docteur Roux, 75015, Paris, France

Corresponding Author: yoann.mouscaz@bioaster.org

BIOSPECIMENS, a web application fostering collaborations in the scope of infectious diseases and microbiology

BIOSPECIMENS [1] is a free to use collaborative platform which brings together project leaders and biological sample holders in the fields of infectious diseases and microbiota. Important changes come with this new major release which included to take some significant choices. Several clinical research engineers, bioinformaticians and informaticians were gathered in the interest of this translational project.

Major technological changes and open source priority

Most of the innovation in the version 1.5 is focused on the technical part of the application which was subjected to a complete overhaul favoring the wide incorporation of open-source components: Symfony [2] as web PHP framework, Doctrine [3] for the ORM (*Object Relational Mapping*) and Twig [4]. Moreover, some notable improvements in the MariaDB [5] database architecture were made to obtain better performances.

BIOSPECIMENS v1.5 is now more hardened to host repositories of data complying with heterogeneous sources and is ready to be proposed as a main portal to academic and industrial collection-holders or to be incorporated as a component of future BIOASTER collaborative projects.

A robust development methodology

Due to project complexity and to the large scale of technical fields covered in the team, we had to find an efficient way to interact between us. Self-hosted and open source DevOps solutions were studied and tested.

Redmine [6] and GitLab [7] platforms appeared to be the good candidates regarding our needs. Concerning the first one, it was used as documentation, Gantt and ticketing tool. In addition, it is usable by each range of users with different level of affinity regarding computer skills; whereas the second one was used by the bioinformaticians and informaticians to develop, to version the source code and to proceed of its continuous integration.

Acknowledgements

Thanks to Amila Malinovic for her works produced on BIOSPECIMENS v1.5 during her Master internship.

References

[1] https://biospecimens.bioaster.org
[2] https://symfony.com/
[3] http://www.doctrine-project.org/
[4] https://twig.sensiolabs.org/
[5] https://www.mariadb.org/
[6] http://www.redmine.org/
[7] https://www.gitlab.com/

Host tropism and host-pathogen interplay of typhoidal *Salmonella enterica*

Ludovic MALLET[1], Claire HOEDE[1], Franck CERUTTI[1], Annick MOISAN[1], Christine GASPIN[1], Isabelle VIRLOGEUX-PAYANT[2], Inna SHOMER[3], Ohad GAL-MOR[3], Thomas SCHIEX[1] and Hélène CHIAPELLO[1]

[1] Inra, UR 875 MIAT, Auzeville, 31326 Castanet-Tolosan, France,[2]Inra, UMR1282 Infectiologie et Santé Publique, F-37380 Nouzilly, France, [3] The Infectious Diseases Research Laboratory, Sheba Medical Center, Tel-Hashomer, 5262100, Israel.

Corresponding Author: helene.chiapello@inra.fr

The species *Salmonella enterica* is one of the most prevalent human and animal pathogens, it includes Non Typhoïdal *Salmonella* (NTS) serovars like Typhimurium and Enteridis, that are generalist pathogens with broad host specificity and Typhoïdal *Salmonella* (TS) serovars, like Typhi and Paratyphi A, that are specialized pathogens strictly adapted to the human host and the cause of an invasive, dangerous disease known as enteric (typhoid) fever [1,2,3].

The SalHostTrop project aims at identifying, characterizing and understanding the human-restricted tropism of Typhoidal *Salmonella* (TS) using comparative dual-RNAseq sequencing and other complementary approaches.

We combine state of the art genome and transcriptome sequencing methods to decipher the molecular basis of host-tropism in clinical strains. We contrast the comparative genomics and differential expression analyses to explore and assess the variability and plasticity of pathogenesis routes among and between typhoidal and non-typhoidal serovars.

We present our on-going work including the Pacbio long-read genomic sequencing, assembly and annotation [4] of a new *S*. Typhi strain (120130191) and the dual RNAseq data analysis of a pilot experiment of *S*. Typhimurium and *S.* Paratyphi A during human epithelial cells infection. The new *S*. Typhi strain includes one circularized complete chromosome and one plasmid of about 4.78 Mb with 4638 coding genes and 106.7 kb with 128 coding genes, respectively. The dual RNAseq pilot first analyses demonstrate the feasibility of the protocol to target both pathogen and host transcripts simultaneously during infection. We also built a *S. enterica* subsp. *enterica* reference phylogenetic tree from the super-alignment of *Salmonella* core genes in 214 complete genomes of various serotypes that is in agreement with previous studies and will be used to explore pseudogene content of serotypes according to their evolutionary history.

Acknowledgements

This work was supported by the Infect-ERA SalHostTrop projet. We thank the platforms GenoToul Bioinfo and GenoToul GetPlage for support, resources and services.

References

[1] Langridge GC, Fookes M, Connor TR, Feltwell T, Feasey N, Parsons BN, Seth-Smith HM, Barquist L, Stedman A, Humphrey T *et al*: **Patterns of genome evolution that have accompanied host adaptation in Salmonella**. *Proc Natl Acad Sci U S A* 2015, **112**(3):863-868.

[2] Gal-Mor O, Boyle EC, Grassl GA: **Same species, different diseases: how and why typhoidal and non-typhoidal Salmonella enterica serovars differ**. *Frontiers in microbiology* 2014, **5**:391.

[3] Grépinet O, Rossignol A, Loux V, Chiapello H, Gendrault A, Gibrat J-F, Velge P, Virlogeux-Payant I: **Genome sequence of the invasive Salmonella enterica subsp. enterica Serotype Enteritidis Strain LA5**. *Journal of Bacteriology* 2012, **194**(9):2387-2388.

[4] Sallet E, Gouzy J, Schiex T: **EuGene-PP: a next-generation automated annotation pipeline for prokaryotic genomes**. *Bioinformatics* 2014, **30**(18):2659-2661.

Towards a new heuristics to compute Consensus Ranking of Big Biological Data

Pierre ANDRIEU[1], Laurent BULTEAU[2], Sarah COHEN-BOULAKIA[1], Alain DENISE[1,3], Anthony LABARRE[2], Adeline PIERROT[1] and Stéphane VIALETTE[2]

[1] Laboratoire de Recherche en Informatique (LRI), Université Paris-Sud, CNRS, Université Paris-Saclay, France

[2] Laboratoire d'Informatique Gaspard-Monge (LIGM), Université Marne-la-Vallée, CNRS, France

[3] Institut de Biologie Intégrative de la Cellule (I2BC), Université Paris-Sud, CNRS, CEA, Université Paris-Saclay, France

Corresponding author: pierre.andrieu@lri.fr, sarah.cohen-boulakia@lri.fr

The aim of biological data ranking is to help users faced with huge amount of data and choose between alternative pieces of information. This is particularly important when querying biological data integration systems, where even very simple queries can return thousands of answers. For instance, searching for the set of human genes involved in breast cancer returns thousands of answers in the reference database EntrezGene without any ranking in terms of importance. The need for ranking solutions, able to order answers, is crucial for helping scientists to organize their time and prioritize the new experiments to be possibly conducted. However, ranking biological data is a difficult task for various reasons: biological data are usually annotation files which reflect expertise, they thus may be associated with various degrees of confidence; the need expressed by scientists may also be taken into consideration whether the most well-known data should be ranked first, or the freshest, etc. As a consequence, although several ranking methods have been proposed in the last years within the bioinformatics community, none of them has been deployed on systems currently in use.

1 Consensus Ranking for Biological data

The approach we propose to follow [1] is to rank biological data by considering two steps. First, several ranking methods are applied to biological data (results are ordered using alternative ranking criteria). Second, we use consensus ranking methods reflecting the input rankings' common points while not putting too much importance on elements classified as "good" by only one or a few rankings. The problem, known as the *median problem* for a set of rankings, is **NP-hard**. However, since providing a consensus ranking is a crucial need for big biological data sets, designing scalable algorithms is highly challenging. Besides, the problem has been mainly studied in the case of *permutations* where elements are strictly ordered while in real applications some elements may be placed at the same position (considered as equally important). The challenge is then to design an algorithm computing one consensus ranking from a set of rankings with ties.

2 Towards a partitioning solution for ranking with ties

We introduce a new algorithm computing a consensus ranking from a set of rankings with ties. The originality of our approach lies in providing an efficient solution (i) based on a graph decomposition of the datasets to partition it efficiently and (ii) having several interesting and fundamental properties, which allow to evaluate the relevance of a given solution and able to provide the exact consensus in many cases. A set of experiments has been conducted on several hundreds of biological and synthetic data sets [2]. First results appear to be very promising, making our algorithm able to compete with the best currently available algorithms while being efficient enough to be used on real settings [3] in particular as the algorithm used on http://conqur-bio.lri.fr/.

Acknowledgements

This work was supported by the CNRS Mastodons *QualiBioConsensus* project.

References

[1] Sarah Cohen Boulakia, Alain Denise, and Sylvie Hamel. Using medians to generate consensus rankings for biological data. In *Scientific and Statistical Database Management - 23rd International Conference, SSDBM 2011, Portland, OR, USA, July 20-22, 2011. Proceedings*, pages 73–90, 2011.

[2] Bryan Brancotte, Bo Yang, Guillaume Blin, Sarah Cohen Boulakia, Alain Denise, and Sylvie Hamel. Rank aggregation with ties: Experiments and analysis. *PVLDB*, 8(11):1202–1213, 2015.

[3] Bryan Brancotte, Bastien Rance, Alain Denise, and Sarah Cohen Boulakia. Conqur-bio: Consensus ranking with query reformulation for biological data. In *Data Integration in the Life Sciences - 10th International Conference, DILS 2014, Lisbon, Portugal, July 17-18, 2014. Proceedings*, pages 128–142, 2014.

Titre : « *Comparaison de pipelines pour la découverte de signatures métagénomiques à partir de séquençage 16S en contexte clinique* »

Auteurs : Aziza Caidi[1], Emma Bergsten[2], Iradj Sobhani[2], Denis Mestivier[1]

[1] Plateforme Bioinformatique/Institut Mondor de Recherche biomédicale (IMRB-Inserm U955)- Université Paris-Est Créteil, Faculté de Médecine de Créteil, 8 rue du Général Sarrail 94010 Créteil cedex, France

[2] Early detection of Colon Cancer using Molecular Markers and Microbiota (EC2M3) – IMRB/EA 7375, Groupe Hospitalier Henri Mondor, Service de Gastroentérologie, 51 Av Maréchal de Lattre de Tassigny, 94010 Créteil, France

Auteur correspondant: aziza.caidi@u-pec.fr

Résumé

Le microbiote intestinal joue un rôle important dans la santé de son hôte et des changements de sa composition (dysbiose) sont associés avec des états pathologiques (obésité, diabète, cancers, etc)[1] [2] .

L'amplification et le séquençage du gène de la sous unité ribosomale 16 (16S) reste la technologie la plus utilisée pour l'identification des microbiotes en métagénomique clinique[3]. Plusieurs pipelines bioinformatiques existent pour permettre l'identification de signatures métagénomiques (compositions différentielles de bactéries) et/ou de bactéries biomarqueurs dans le diagnostic[4].

Dans le contexte clinique du cancer colo-rectal (CCR) qui est l'un des trois cancers les plus fréquents, des études ont mis en évidence l'implication de quelques bactéries dans le développement de ce cancer[4]. Ces signatures métagénomiques ont été identifiées par des pipelines différents (Mothur, QIIME, MEGAN), mais aucune comparaison de ces pipelines n'a été effectuée, rendant difficile la confrontation des différentes signatures métagénomiques publiées.

L'objectif de ce travail, qui bénéficie de deux cohortes monocentriques d'une même population, est de préciser les limites et la reproductibilité de ces pipelines dans l'identification de signatures métagénomiques.

La signature métagénomique sera évaluée par le package metagenomeSeq du logiciel R pour chacune des tables d'annotation produites par les pipelines sur chaque cohorte. La reproductibilité des pipelines sera présentée pour les différents niveaux taxonomiques et des règles pourront être proposées de choix d'un pipeline dans ce contexte clinique.

Références

[1] C. Jin, J. Henao-Mejia, and R. A. Flavell, "Innate immune receptors: Key regulators of metabolic disease progression," *Cell Metab.*, vol. 17, no. 6, pp. 873–882, 2013.

[2] I. Sobhani *et al.*, "Microbial dysbiosis in colorectal cancer (CRC) patients," *PLoS One*, vol. 6, no. 1, 2011.

[3] R. S. Mandal, S. Saha, and S. Das, "Metagenomic Surveys of Gut Microbiota," *Genomics, Proteomics Bioinforma.*, vol. 13, no. 3, pp. 148–158, 2015.

[4] G. Zeller *et al.*, "Potential of fecal microbiota for early-stage detection of colorectal cancer.," *Mol. Syst. Biol.*, vol. 10, no. 11, p. 766, 2014.

Horizontal gene transfer from viruses in the genomes of plant-parasitic nematodes

Carole BELLIARDO, Corinne RANCUREL, Etienne G.J. DANCHIN and Marc BAILLY-BECHET
Institut Sophia Agrobiotech, INRA, CNRS, Univ. Côte d'Azur, 06900 Sophia Antipolis, France

Corresponding author: carole.Belliardo@etu.unice.fr

1 Introduction

Root-knot nematodes, genus *Meloidogyne*, are one of the most damaging plant-pest around the world and sufficient control methods are missing. To develop new control methods, we need to better understand the biological processes and evolutionary history of *Meloidogyne*. For all plant-parasitic nematodes characterized at the omics level, several horizontal gene transfer (HGT) from bacteria or fungi have been highlighted. HGT can originate from soil dwelling organisms and pathogens of nematodes or plants [1]. The role of viruses as potential donors is unclear, and few nematode viruses are known.

2 Objectives

Recently, metagenomic sequencing has increased the availability of virus sequences in public databases[2]. These data open a new opportunity to determine whether there is a viral contribution to the *Meloidogyne* genomes. Here, we aim at identifying *Meloidogyne* genes of viral origins and the families of potential viral donors.

3 Material and Methods

We used two approaches to identify and characterize candidate HGT of viral origins. First, we detect HGT based on sequence homology with two different tools: (i) Alienness, a taxonomy aware BLAST of *Meloidogyne* proteins against nrNCBI-DB and calculation of difference in magnitude of e-value between metazoan and viral hits[3]; (ii) Retrieval of viral HMM profiles from ImgVR-DB [2] and scan against *Meloidogyne* proteins. In second step, we look at GC percent and codon usage bias of *Meloidogyne* protein coding genes by multivariate analysis. This way, we compare the composition of genes acquired from viruses to the rest of the genome, to determine the molecular signatures of viral HGT and discriminate old and recent HGT.

4 Results and discussion

We found 80 sequences with putative viral origin in coding sequences of the model root-knot nematode, *Meloidogyne incognita*. Among those, there is a significant enrichment of *Herpesviridae*, *Baculoviridae* and *Adenoviridae*, animal virus strains. All of these genes are supported by expression data, so we suppose a domestication by host genome. Analysis of the codon usage bias brings to light a significant difference between these viral sequences: on one side, domesticated sequences indicating old HGT, with *Baculoviridae* lineage annotations, and on the other side recent HGT with *Adenoviridae* and *Herpesviridae* annotation. Comparative studies with other *Meloidogyne* and *C. elegans*, converge toward that a part of HGT is ancient and the other part is recent.
These results provide a field of investigation to further characterize the candidate viral donors at the species level that could today infect *Meloidogyne* species. These findings could be useful for new bio-controls methods or in biotechnology. Indeed, no virus based bio-control or transformation methods are known in *Meloidogyne* so far.

References

[1] E. G. J. Danchin & al. Multiple lateral gene transfers and duplications have promoted plant parasitism ability in nematodes. *Proceedings of the National Academy of Sciences*, October 2010.

[2] David Paez-Espino, Emiley A. Eloe-Fadrosh, Georgios A. Pavlopoulos, Alex D. Thomas, Marcel Huntemann, Natalia Mikhailova, Edward Rubin, Natalia N. Ivanova, and Nikos C. Kyrpides. Uncovering Earth's virome. *Nature*, 536(7617):425–430, August 2016.

[3] C. Rancurel & al. Rapid detection of horizontal gene transfers in metazoan genomes. https://f1000research.com/posters/1096828, November 2014.

Mutation of Tyr137 of the universal Escherichia coli fimbrial adhesin FimH relaxes the tyrosine gate prior to mannose binding

Eva-Maria Krammer[1], Goedele Roos[1], Martine Prévost [2], Julie Bouckaert and Marc F Lensink[1]

[1] University of Lille, CNRS UMR8576 UGSF, F-59000 Lille, France
[2] Structure et Fonction des Membranes Biologiques, Université Libre de Bruxelles (ULB), Brussels, Belgium

Corresponding Author: eva-maria.krammer@univ-lille1.fr

The most prevalent diseases manifested by Escherichia coli are acute and recurrent bladder infections and chronic inflammatory bowel diseases such as Crohn's disease [1,2]. E. coli clinical isolates express the FimH adhesin, which consists of a mannose-specific lectin domain connected via a pilin domain to the tip of type 1 pili. Although the isolated FimH lectin domain has affinities in the nanomolar range for all high-mannosidic glycans, differentiation between these glycans is based on their capacity to form predominantly hydrophobic interactions within the tyrosine gate at the entrance to the binding pocket [3].

Single-residue mutations in the tyrosine gate (Tyr48Ala and Tyr137Ala) have a moderate effect on the affinity of FimH for mannose, but a markedly lower affinity is observed for heptyl- and biphenyl-substituted mannosides that intercalate in the tyrosine gate. In the crystals of the Y137A mutant, a breakdown of the binding site with a severe loss of specificity is observed [4]. Using quantum-mechanical calculations, we could demonstrate that the wild-type Tyr137 introduces strain in the polypeptide backbone. This maintains a high energetic potential that is normally only released upon the binding of an oligomannosidic ligand. Using molecular-dynamics simulations, we could highlight that this energetic potential is coupled to the other tyrosine of the gate, Tyr48, via the inner mannose-binding residue Ile52. In conclusion, the mutation of Tyr137 to alanine relaxes the binding site prematurely, whereby the stringent selectivity of the FimH lectin for mannose is disrupted and the binding affinity decreases.

References

[1] Goneau et al. *Subinhibitory Antibiotic Therapy Alters Recurrent Urinary Tract Infection Pathogenesis through Modulation of Bacterial Virulence and Host Immunity* mBio, 2015. 6: p. e00356-15.
[2] Darfeuille-Michaud et al. *High prevalence of adherent-invasive Escherichia coli associated with ileal mucosa in Crohn's disease*. Gastroentrology 2004. **127:** p 412-421.
[3] Roos et al. *Validation of Reactivity Descriptors to Assess the Aromatic Stacking within the Tyrosine Gate of FimH.* ACS Med Chem Lett. 2013. **4**: p.1085-1090.
[4] Rabbani et al, *Mutation of Tyr137 of the universal Escherichia coli fimbrial adhesin FimH relaxes the tyrosine gate prior to mannose binding,* IUCrJ, 2017. **4**: p.7-23

Collapsing reads while maintaining qualities : srnaCollapser

Walid BEN SAOUD BENJERRI, Christine GASPIN and Matthias ZYTNICKI
MIAT, 24 Chemin de Borde Rouge, 31320, Auzeville, France

Corresponding author: matthias.zytnicki@inra.fr

1 Background

sRNAs are a class of non coding RNA molecules. Many of them play key roles in the cell life, such as microRNAs, that target messenger RNA to inhibit translation. Besides their short size ∼15–200, they are also characterized by their diversity of abundance as some of them are highly transcribed [1]. Sequencing technologies are used to get the ordered list of nucleotides of many sRNAs in a sample. The output takes the form of a FASTQ file. However, sequencing is error prone, there is no guarantee that the correct nucleotide matches the one output by the machine. To estimate the reliability of a base calling, current technologies attach to each nucleotide a quality value which is function of the probability that the correct base was called.

The amount of generated data is huge and highly redundant. Because of that, RNA sequencing data analysis is preceded by a preprocessing step that aims to remove this redundancy. Typically, users will eliminate duplicated reads keeping only one sequence while maintaining the number of occurrences for this sequence, discarding the associated read original qualities. This is known by "collapsing" [2].

2 A new approach for small RNA read collapsing

We propose an alternative way of collapsing, wherein a synthetic quality is calculated for all the duplicates of the same sequence. Keeping qualities, as opposed to the naive usage, will give additional information at the next step of the pipeline, the read mapping. It is especially useful for reads having few duplicates or unique reads. The emphasis is put on efficiency of the implementation so that it can be used as part of larger pipelines.

The synthetic quality is calculated as follows. For each base b, the synthetic quality value is given by the maximum over the qualities associated to b in the set of identical reads to collapse.

Two other functionalities are provided. For each sequence, its number of occurrences is computed by sample. Finally results are given in a sorted manner as a FASTQ-like file.

3 Implementation

To make the approach efficient, reads are inserted in a Trie. Usage of a trie allows to aggregate similar sequences to minimize spatial disparity which offers quick retrieval (or quick rejection) when inserting new sequences. High speed is achieved by using low level optimizations such as bit-manipulations.

4 Performance Evaluation

We compared the runtime and space usage of our approach to a naive solution built-on the Unix tool sort (with an additional script to collapse qualities), and a hashtable solution. Result show performance improvement of srnaCollapser over the other approaches.

Method	1 file (1.7 GB)	3 files (4.8 GB)	6 files (9.3 GB)	9 files (14.1 GB)
srnaCollapser	36s	1m28s	3m42s	5m34s
GNU Sort	4m	11m50s	$\geq$ 20m	out of memory
Hashtables	22s	1m30s	4m20s	out of memory

References

[1] Shirley Tam, Ming-Sound Tsao, and John D. McPherson. Optimization of miRNA-seq data preprocessing. *Briefings in Bioinformatics*, 16(6):950, 2015.

[2] Filipe Borges and Robert A. Martienssen. The expanding world of small RNAs in plants. *Nat Rev Mol Cell Biol*, 16(12):727–741, 2015.

Posters partenaires

BioMAn™: a user-friendly interface for targeted metagenomic data visualization and analysis

Pauline Vaissié[1], Christophe Camus[1], Yao Amouzou[1], Thomas Carton[1], Sophie Le Fresne-Languille[1], Françoise Le Vacon[1], Murielle Cazaubiel[1] and Sébastien Leuillet[1]

[1] Biofortis Mérieux NutriSciences

Corresponding Author: pauline.vaissie@mxns.com

With the recent advances in the field of next-generation sequencing (NGS), metagenomics allow to explore the biodiversity of microbial ecosystems or microbiota. Dedicated bioinformatic pipeline focusing on targeted metagenomics (16S rRNA) provides to biologists the bacterial composition in OTUs (Operational Taxonomic Units) of the samples. Nevertheless, NGS produces massive data which requires substantial computer processing to extract information. Faced with this large amount of data, their visualization and appropriate statistical analysis are essential for scientists to adequately explore and interpret their experiments. In this context, we have developed an R [1] and Shiny [2] web based platform called BioMAn™ (Biofortis Metagenomics Analysis) which mixes the statistical power of dedicated R packages (metagenomeSeq, mixOmics…) with a user friendly web design. This interface allows users to interactively look into their project by manipulating, filtering or gathering information for further interpretation or communications purposes. Focusing on a subgroup of samples is made very easy by the integration of metadata table (information on samples such as experimental conditions). The core of the tool is focused on data visualization, which offers the possibility to depict taxonomic composition throughout several graphical interactive representations such as barplots, boxplots, heatmaps, Krona [3] or hierarchical trees. BioMAn™ also provides information (tables and graphs) about diversity indices to help users in the interpretation of results. This turnkey product is an easy way for scientists to conduct ordination analysis such as PCoA with a lot of customizable graphical and analytical options. The platform can also be used to run specific statistical analysis like discriminant analysis (LDA and FDA). Other statistical approaches are currently being added to the application (PERMANOVA/ANOSIM, differential analysis…) in order to create the fullest possible metagenomic toolbox. During the process, user can easily retrieve objects by downloading them in high quality or by inserting them one by one into a custom PowerPoint template. BioMAn™ is deployed on a Shiny Server Pro, implemented by a secure health data hosting provider according to the French regulatory requirements, to protect the confidentiality, integrity and availability of patient and user data.

References

[1] R Core Team. R:A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria, 2013.

[2] Winston Chang, Joe Cheng, JJ Allaire, Yihui Xie and Jonathan McPherson (2017). shiny: Web Application Framework for R. R package version 1.0.0. https://CRAN.R-project.org/package=shiny

[3] Ondov, Brian D., Nicholas H. Bergman, and Adam M. Phillippy. "Interactive metagenomic visualization in a Web browser." *BMC bioinformatics* 12.1 (2011): 385.

WHORMSS : un nouvel outil pour l'exploration taxonomique et fonctionnelle sans à priori des métagénomes.

Yannick LAURENT[1], Jérémie DENONFOUX[2], Antoine BODEIN[2], Stéphanie FERREIRA[2]

[1] Genoscreen, Service Bioinformatique, 1 rue du Professeur Calmette, 59000 Lille, France

[2] Genoscreen, Service Recherche-Développement-Innovation, Communautés Microbiennes, 1 rue du Professeur Calmette, 59000 Lille, France

Contact : yannick.laurent@genoscreen.fr

Les nouvelles technologies de séquençage (NGS) permettent d'accéder, à très haut débit, à des informations génomiques complexes et conséquentes. Cependant, la taille encore trop courte des lectures, les erreurs de séquençage, la quantité de données à analyser ou encore la mixité variable des matrices initiales nécessite le développement d'outils bioinformatiques adaptés et performants permettant d'extraire les informations microbiennes au sein des métagénomes. Les méthodes sans a priori de type Whole Metagenome Shotgun (WMS) s'affranchissant de la PCR et se basant sur un séquençage direct de l'ensemble des génomes présents au sein d'un échantillon, elles représentent des stratégies d'interrogations originales et puissantes des environnements. Ces dernières permettent d'avoir accès notamment à la fraction inconnue et non cultivable des communautés microbiennes, représentant un vivier de nouvelles fonctions. Cependant, l'importante diversité de ces communautés, la connaissance partielle que nous en avons à l'heure actuelle combinées à l'information fragmentaire obtenue et la courte taille des lectures rendent l'annotation des séquences complexe. Les approches métagénomiques à haut-débit WMS représentent donc un réel défi analytique.

Afin de pouvoir proposer un outil automatisé d'analyse de données métagénomiques à haut-débit de type WMS dans un contexte d'étude des communautés microbiennes, le pipeline WHORMSS pour WHOle Recovering from Metagenome Shotgun Sequencing a été développé. Ce dernier est dédié à l'analyse de données de séquençage de type Illumina « paired-end ». WHORMSS intègre un contrôle qualité des séquences, une étape d'assemblage et une stratégie d'annotation taxonomique et fonctionnelle, répondant aux contraintes des analyses métagénomiques WMS à haut-débit. De plus, un mode de recherche ciblée de fonctions microbiennes permet de mettre en lumière des capacités métaboliques connues et inconnues. Les performances de WHORMSS en termes de qualité des affiliations mais également de rapidité et de potentiel d'étude de la structure et fonctions des communautés, ont été éprouvées sur un échantillon de composition maîtrisée et sur des échantillons complexes environnementaux (eau de mer et microbiote intestinal).

WHORMSS a démontré tout son potentiel pour décrire les communautés microbiennes dans un cadre d'étude d'échantillons environnementaux complexes par des approches sans à priori. WHORMSS est également capable de mettre en lumière et à façon des fonctions métaboliques d'intérêt.

Mots clés : Métagénomique sans a priori, NGS, diversité microbienne, pipeline, annotation taxonomique et fonctionnelle

CRISPR LifePipe®: tools for the design of gRNAs and donor sequence required for genome editing using CRISPR/Cas9 system

Virginie CHESNAIS[1], Emmanuel CHAPLAIS[1], Alban OTT[1] and Eric GINOUX[1]

[1] Life&Soft, 8b avenue Descartes, 92350, Plessis-Robinson, France

Corresponding Author: vchesnais@lifeandsoft.com

The introduction of targeted genomic sequences modifications by CRISPR technology into living cells is becoming a powerful tool for gene therapy or disease modelling [1]. CRISPR only requires a nuclease and customized nucleic sequences. Preliminary bioinformatics analysis for both gRNA design and donor template can improve the success of the experiment. This is where CRISPR LifePipe will make genome editing as simple as using a text editor.

The gRNA (guide RNA) is a short RNA sequence which guides the Cas9 endonuclease to the targeted region to cut on the genome. gRNA is crucial for CRISPR gene editing, because it provides targeting efficiency and specificity on the region of interest on the genome, while limiting the non-specific off-targets. Our gRNA design tool is built to allow you to target different genomic regions: (i) targeting an exon present on most of the transcripts of a gene, (ii) targeting an exon or an intron of a transcript, (iii) targeting UTR region, (iv) targeting a particular amino acid or (v) targeting a particular DNA sequence. The efficiency of gRNAs is determined by various annotations like secondary structure, presence of SNP on the sequence, prediction score… The research of off-targets for all gRNAs assesses their specificity and help the user to choose the best gRNA.

The donor template is a DNA sequence inserted into the cell along with the gRNA and the Cas9 endonuclease to replace DNA sequence of the cell. Different donor sequences can be designed with our tool according to the modification that is expected in the cell: (i) insertion of a mutation, (ii) gene tagging, (iii) gene knock-out and/or (iv) insertion of a DNA fragment, like a selection cassette. During the process, the donor sequence is inactivated to prevent the Cas9 to cut the donor sequence. Quality control and annotation are also performed to assess the quality of the donor.

Our CRISPR LifePipe is based on a workflow built with the snakemake tool [2]. An intuitive, responsive and ergonomic web interface make a better user experience. One-page web application was developed with Django and Angular UI [3]. Finally, all source code and third-party applications were encapsulated in a docker image, which provides an easier deployment in any production informatics structure [4].

In summary, CRISPR LifePipe tools have been developed to meet all the needs of the CRISPR users. These user-friendly tools will facilitate and improve all steps required for a high quality and successful CRISPR experiment preparation.

References

[1] L. Cong, F.A. Ran, D. Cox, S. Lin, R. Barretto, N. Habib, P.D. Hsu, X. Wu, W. Jiang, L.A. Marraffini, F. Zhang, Multiplex genome engineering using CRISPR/Cas systems, Science. 339 (2013) 819–823. doi:10.1126/science.1231143.

[2] J. Köster, S. Rahmann, Snakemake—a scalable bioinformatics workflow engine, Bioinformatics. 28 (2012) 2520–2522. doi:10.1093/bioinformatics/bts480.

[3] AngularUI for AngularJS, (n.d.). https://angular-ui.github.io/ (accessed March 20, 2017).

[4] Docker, Docker. (n.d.). https://www.docker.com/ (accessed March 20, 2017).

Bioinfo-fr.net : présentation du blog communautaire scientifique francophone par les Geekus biologicus

LA COMMUNAUTÉ DE GEEKUS BIOLOGICUS[1]

1 Bioinfo-fr.net

Corresponding Author: admin@bioinfo-fr.net

Pourquoi ?

"Bioinformatique ? C'est quoi ça ? De l'informatique respectueuse des terres ???". Qui n'a jamais eu ce genre de remarque sur notre profession ? Qui n'a jamais peiné pour expliquer à son entourage son travail de tous les jours ? Depuis quelques années, une petite communauté de bioinformaticien·ne·s francophones s'est formée suite à un simple constat : il y avait bel et bien un trou dans l'Internet francophone par rapport à notre science ! Nous ne pouvions pas laisser cela tel quel, c'était devenu notre mission. www.bioinfo-fr.net était né !

Qui ?

Au commencement, nous n'étions qu'une petite dizaine avec tout plein d'idées par dessus la tête et une envie commune de faire avancer les choses. Aujourd'hui la recette a séduit, les fondations ont été posées, le mécanisme est bien huilé et nous sommes plus d'une soixantaine de collaborateurs bénévoles dispersés de partout à travers le monde, toujours avec la même motivation. Pas de hiérarchie stricte clairement mise en place, si ce n'est un petit groupe d'administrateurs dont les rôles principaux sont de maintenir le site à jour, veiller à la bonne cohérence des articles, fournir un planning de publication, accueillir les nouveaux venus, relancer les auteurs et les relecteurs de temps en temps, communiquer avec l'extérieur et essayer de mettre en place les nouvelles idées venant de tout un chacun.

Comment ?

D'abord via un canal IRC (#bioinfo-fr, réseau FREENODE). Mais très vite nous avons constaté que de nombreuses questions similaires ressortaient et qu'il serait plus efficient de garder les réponses sur un support écrit et de façon pérenne. Le blog nous a semblé être le support le mieux adapté cela. Chaque semaine nous nous efforçons de fournir un article qui a subit un processus de relecture scientifique robuste et qui doit avoir un rapport de près ou de loin avec la bioinformatique. Un article doit entre-autre chose permettre au lecteur, averti ou non, de comprendre une méthode, de reproduire une expérience, de plonger directement dans un code solutionnant un problème biologique ou encore de l'informer sur une toute nouvelle découverte. Au fil des années et des articles nous avons essayé de décomposer ces articles en catégories afin de faciliter la visite et la recherche du lecteur.

Bilan

En quelques années nous avons réussi à tisser un large réseau de professionnels/étudiants/passionnés capable de produire des articles d'intérêt public et de répondre à des problématiques axées autour de la bioinformatique. Cela nous a permis également de mieux connaitre nos spécialités, nos manières de vivre, notre métier, et de représenter à notre niveau la force de la bioinformatique francophone. Nous ne comptons pas en rester là et vous encourageons fortement à venir discuter avec nous les Geekus biologicus. Peut-être, qui sait, franchirez-vous le cap et nous rejoindrez-vous dans l'aventure !

Liste des contributeurs

Journées Ouvertes en Biologie, Informatique et Mathématiques

Lille, 3-6 juillet 2017