Proceedings

Posters
Démos

Preface

Welcome, Bienvenue, Degemert mat to the 520 attendees of the 23rd JOBIM edition!

What a nice periodicity! Exactly twenty years after the first JOBIM edition organized in Brittany at St Malo, and ten years after the last edition organized in Rennes, we are happy to welcome once again the bioinformatics community in Rennes.

After two years of remote JOBIM conferences at Montpellier and Paris, the bioinformatic community has fully taken the opportunity to meet physically in Rennes. The number of participants exceeded expectations even before the end of early-bird registration! Accordingly, the Program Committee received an impressive number of 298 submissions and selected among them 42 oral contributions, 226 posters and 17 demos. We take this opportunity to acknowledge and express our gratitude to all members of the Program Committee for their crucial reviewing of all these contributions.

We sincerely thank our six keynote speakers who accepted to prepare exciting talks for this JOBIM edition: Cédric Notredame, Pierre Peterlongo, Hélène Morlon, Anne-Laure Boulesteix, Guillaume Bourque and Raphaël Guérois. We look forward to their presentations!

As in previous editions, half a day during JOBIM is dedicated to specialized mini-symposiums. This year, we are welcoming 5 mini-symposiums covering a wide range of bioinformatics topics, and we would like to take this opportunity to thank their organisers.

In the end, we are grateful to the institutional supports of the SFBI, GDR BIM, IFB and warmly thank all the members of the Organising Committee (with a special thanks to Clara, Edith, Fabrice, Jeanne, Marie and Stéphanie) who made it possible to hold JOBIM in Rennes this year.

Claire Lemaitre, Emmanuelle Becker and Thomas Derrien

Program committee

Emmanuelle Becker (Université de Rennes)
Thomas Derrien (CNRS)
Claire Lemaitre (Inria)

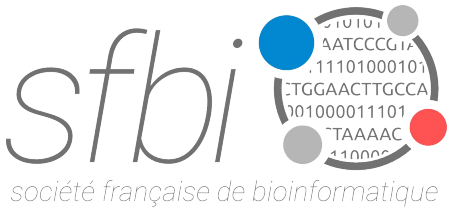
Sophie Abby (CNRS)
Julie Aubert (INRAE)
Benoît Ballester (INSERM)
Anaïs Baudot (CNRS)
Sèverine Bérard (Université de Montpellier)
Yuna Blum (CNRS)
Jérémie Bourdon (CNRS)
Christine Brun (CNRS)
Anne-Claude Camproux (Université Paris-Diderot)
Frédéric Cazals (Inria)
Isaure Chauvot De Beauchêne (Inria)
Hélène Chiapello (INRAE, représentant l'IFB)
Rayan Chikhi (CNRS)
Olivier Collin (CNRS, représentant l'IFB)
Erwan Corre (CNRS, représentant la SFBI)
Olivier Dameron (Université de Rennes)
Sarah Djebali (INSERM)
Patrick Durand (IFREMER)
Damien Eveillard (CNRS)
Anna-Sophie Fiston-Lavier (Université de Montpellier, représentant la SFBI)
Emmanuel Giudice (Université de Rennes)
Gilles Labesse (CNRS)
Vincent Lacroix (Université de Lyon)
Sandrine Lagarrigue (INRAE)
Aurélie Lardenois (Université de Rennes)
Yann Le Cunff (Université de Rennes)
Charles Lecellier (CNRS)
Emmanuelle Lerat (CNRS)
Camille Marchet (CNRS)
Mahendra Mariadassou (INRAE, représentant le GDR BIM)
Pierre Neuvial (CNRS)
Anna Niaraki (Université Paris-Saclay)
Jacques Pécreaux (CNRS)
Eric Pelletier (CEA)
Pierre Peterlongo (Inria)
Yann Ponty (CNRS)
Eric Rivals (CNRS, représentant le GDR BIM)
Hugues Roest Crollius (CNRS)
Mikaël Salson (Université de Lille)
Céline Scornavacca (CNRS)
Patricia Thébault (Université de Bordeaux)
Nathalie Théret (INSERM)
Raluca Uricaru (Université de Bordeaux)
Nathalie Vialaneix (INRAE)

Organizing committee

Edith Blin (Inria)
Marie Le Roïc (Inria)
Fabrice Legeai (INRAE)
Stéphanie Robin (INRAE)

Jacky Ame (INRAE/Inria)
Moana Aulagner (Inria)
Catherine Belleannée (Univ. Rennes 1)
Cécile Beust (Inria)
Anthony Bretaudeau (INRAE)
Karel Brinda (Inria)
Matéo Boudet (INRAE)
Matthieu Bouguéon (INSERM)
Olivier Boullé (Inria)
Konogan Bourhy (CNRS)
Nicolas Buton (Université Rennes 1)
Thomas Chaussepied (CNRS)
Guillaume Collet (Université Rennes 1)
Olivier Collin (CNRS)
François Coste (Inria)
Olivier Dameron (Université Rennes 1)
Clara Delahaye (Université Rennes 1)
Olivier Dennler (Université Rennes 1)
Siegfried Dubois (Inria)
Victor Epain (Inria)
Roland Faure (Université Rennes 1)
Ludovic Fourteau (INRAE)
Kévin Gazengel (INRAE)
Jeanne Got (CNRS)
Garance Gourdel (ENS/Inria)

Anne Guichard (Inria)
Khodor Hannoush (Inria)
Christophe Héligon (CNRS)
Gaëtan Hervé (Université Rennes 1)
Camille Juigné (INRAE/IRISA)
Camille Kergal (Université Rennes 1)
Dominique Lavenier (CNRS)
Téo Lemane (Inria)
Yann Le Cunff (Université Rennes1)
Matthias Lorthiois (CNRS)
Stéphanie Mottier (CNRS)
Jacques Nicolas (Inria)
Laurence Noël (INSERM)
Pierre Peterlongo (Inria)
Thomas Picouet (CNRS)
Lucas Robidou (Inria)
Sandra Romain (Inria)
Emeline Roux (Université Rennes1)
Baptiste Ruiz (INRAE/Inria)
Olivier Sallou (Université Rennes 1)
Nathalie Théret (INSERM)
Florian Thonier (Inria)
Kerian Thuillier (CNRS)
Yael Tirlet (CNRS)


sfbi
société française de bioinformatique


cnrs
GDR
Groupement
de recherche
BIM Bioinformatique moléculaire


IFB
INSTITUT FRANÇAIS DE BIOINFORMATIQUE


INRAE


Inria


GDR
GÉNOMIQUE
ENVIRONNEMENTALE


UNIVERSITÉ DE
RENNES 1


Oxford
NANOPORE
Technologies


illumina


Hewlett Packard
Enterprise


Biogenouest


Ville
Métropole


FONDATION
RENNES 1
Progresser Innover Entreprendre

Table des matières

Démos

RiboTaxa: Combined approaches for taxonomic resolution down to the species level from metagenomics data revealing novelties

Oshma CHAKOORY[1], Sophie COMTET-MARRE[1], Pierre PEYRET[1]

[1]Université Clermont Auvergne, INRAE, MEDIS, F-63000 CLERMONT-FERRAND, France

Corresponding author: pierre.peyret@uca.fr

Metagenomic classifiers are widely used for the taxonomic profiling of metagenomic data and estimation of taxa relative abundance. Small subunit rRNA genes are nowadays a gold standard for phylogenetic resolution of complex microbial communities, although the power of this marker come down to its use as full-length. We benchmarked the performance and accuracy of rRNA-specialized versus general-purpose read mappers, reference-targeted assemblers and taxonomic classifiers. We then built a pipeline called RiboTaxa to generate a highly sensitive and specific metataxonomic approach. Using metagenomics data, RiboTaxa gave the best results compared to other tools (Kraken2, Centrifuge (1), METAXA2 (2), PhyloFlash (3)) with precise taxonomic identification and relative abundance description giving no false positive detection. Using real datasets from various environments (ocean, soil, human gut) and from different approaches (metagenomics and gene capture by hybridization), RiboTaxa revealed microbial novelties not seen by current bioinformatics analysis opening new biological perspectives in human and environmental health.

In a study focused on corals' health involving 20 metagenomic samples (4), affiliation of prokaryotes was limited to the family level with *Endozoicomonadaceae* characterising healthy octocoral tissue. RiboTaxa highlighted 2 species of *uncultured Endozoicomonas* which were dominant in the healthy tissue. Both species belonged to new genus opening new research perspectives on corals' health.

Applied to metagenomics data from a study on human gut and extreme longevity (5), RiboTaxa detected the presence of an uncultured archaeon in semi-supercentenarians (aged 105 to 109 years) highlighting a new archaeal genus not yet described and 3 new species belonging to the *Enorma* genus that could be species of interest participating in the longevity process.

RiboTaxa is user-friendly, rapid, allowing microbiota structure description from any environment and the results can be easily interpreted. This software is freely available at https://github.com/oschakoory/RiboTaxa under the GNU Affero General Public License 3.0.

References

1. Kim,D., *et al*. (2016) Centrifuge: rapid and sensitive classification of metagenomic sequences. *Genome Res.*, **26**, 1721.

2. Bengtsson-Palme,J., *et al*. (2015) METAXA2: improved identification and taxonomic classification of small and large subunit rRNA in metagenomic data. *Mol. Ecol. Resour.*, **15**, 1403–1414.

3. Gruber-Vodicka,H.R., *et al*. (2019) phyloFlash – Rapid SSU rRNA profiling and targeted assembly from metagenomes. *bioRxiv*, 10.1101/521922.

4. Keller-Costa,T.,*et al*. (2021) Metagenomic insights into the taxonomy, function, and dysbiosis of prokaryotic communities in octocorals. *Microbiome*, **9**, 72.

5. Rampelli,S., *et al.* (2020) Shotgun Metagenomics of Gut Microbiota in Humans with up to Extreme Longevity and the Increasing Role of Xenobiotic Degradation. *mSystems*, **5**, e00124-20.

Computing lowest common ancestors on SAM files with sam2lca

Maxime BORRY[1], Alexander HÜBNER[1,2] and Christina WARINNER[1,2,3]

[1] Microbiome Sciences Group, Max Planck Institute for Evolutionary Anthropology, Department of Archaeogenetics, Leipzig, Germany

[2] Department of Paleobiotechnology, Leibniz Institute for Natural Product Research and Infection Biology - Hans Knöll Institute (HKI), Jena, Germany

[3] Faculty of Biological Sciences, Friedrich-Schiller Universität Jena, Jena, Germany

[4] Department of Anthropology, Harvard University, Cambridge, MA, United States of America

Corresponding Author: maxime_borry@eva.mpg.de

In a typical shotgun metagenomics approach, after the DNA of an ecological community has been sequenced, it is compared to a genetic reference database of organisms with known taxonomy.
Even though the number of DNA sequences and genomes in reference databases is constantly growing, there are still instances where a query sequence will not have a direct match in a reference database, but it will instead align to one or more distantly related reference organisms.
To resolve this ambiguity, one method is to place the query sequence higher in the taxonomic tree at the common ancestor of all ambiguously matched reference sequences, using an algorithm known as lowest common ancestor (LCA).
While there are many different metagenomics classifiers and taxonomic profilers currently available, there is still a need for a program that can perform LCA from the common SAM alignment file format.
Here, we present sam2lca, a program to perform LCA from a SAM/BAM/CRAM file.
Because sam2lca uses the common SAM alignment file format as input, it is easy to use in combination with any SAM-producing alignment program applied to a metagenomics dataset.
Furthermore, with its command line interface, python API, and containerization thanks to bioconda, it is easy to integrate within already existing metagenomics processing pipelines.

Thirdkind : Drawing phylogenetic encounters up to 3 reconciliation levels.

Simon Penel[1], Hugo Menet[1], Théo Tricou[1], Vincent Daubin[1] and Eric Tannier[1,2]

[1] Laboratoire de Biométrie et Biologie Evolutive, UMR5558 CNRS/UCBL, 69622 Villeurbanne,France

[2] BEAGLE Centre de Recherche Inria Lyon,69622 Villeurbanne,France

Corresponding Author: simon.penel@univ-lyon1.fr

Abstract

The history of a species is closely related to the history of its genes. Connecting the evolution of a genome to the evolution of its genes is a way to describe this relationship. In this context, reconciliation of the genes with the species consists into mapping the nodes of a gene tree and the associated events (speciation, duplication, loss, tranfer) to the nodes of the species tree. Reconciliation can as well be used to map the history of a parasite with the history of a host, or to map the history of a protein domain with the history of a sequence.

Visualisation of phylogenetic reconciliations are proposed by various programs and interfaces as NOTUNG [1], SylvX [2], Treerecs [3], Jane [4], eMPRess [5] and Capybara [6]. However at the exception of SylvX, all are integrated in a specific reconciliation software and cannot visualise reconciliations produced by others. None of these software is handling recPhyloXML [7], a XML format proposed as a standard to describe phylogenetic reconciliations, and none of them is generic to any kind of reconciliation (for example SylvX does not allow temporary free living symbionts, as it is not allowed for genes to live outside a genome) nor can handle multiple horizontal transfer (i.e. several genes transfered with the same donor and recipient) and the consideration of numerous possible scenarios. DoubleRecViz [8] uses a derived version of recPhyloXML, adding a transcript level to gene and species format but without support for horizontal transfers.

Eventually there is no software able to combine two nested reconciliations i.e. to get in a single representation the gene/symbiont reconciliation and the symbiont/host reconciliation.

Here we present Thirdkind [9] a very simple command-line software allowing the user to easily generate graphical output (svg) from one or several recphyloXML files with a large choice of options (as for example orientation, police size, branch length, multiple gene trees, multiples species trees, multiple files, redundant transfers handling, etc.) and to handle the display of two nested reconciliations (displaying a gene/symbiont/host reconciliation for example).

Home page : https://github.com/simonpenel/thirdkind/wiki

References

[1] K Chen *et al.* NOTUNG: a program for dating gene duplications and optimizing gene family trees. *J. Comput. Biol.*, (7):429–447, 2010.

[2] F Chevenet *et al.* SylvX: a viewer for phylogenetic tree reconciliations. *Bioinformatics*, (32):608–610, 2016.

[3] N Comte *et al.* Treerecs: an integrated phylogenetic tool, from sequences to reconciliations. *Bioinformatics*, (36):4822–4824, 2020.

[4] C Conow *et al.* Jane: a new tool for the cophylogeny reconstruction problem. *Algorithms Mol Biol.*, DOI: 10.1186/1748-7188-5-16, 2010.

[5] S Santichaivekin *et al.* eMPRess: a systematic cophylogeny reconciliation tool. *Bioinformatics*, (37):2481–2482, 2021.

[6] Y Wang *et al.* Capybara: equivalence ClAss enumeration of coPhylogenY event-BAsed ReconciliAtions. *Bioinformatics*, (36):4197–4199, 2021.

[7] W Duchemin *et al.* RecPhyloXML: a format for reconciled gene trees. *Bioinformatics*, (34):3646–3652, 2018.

[8] E Kuitche *et al.* DoubleRecViz: a web-based tool for visualizing transcript–gene–species tree reconciliation. *Bioinformatics*, (37):1920–1922, 2021.

[9] S Penel *et al.* Thirdkind: displaying phylogenetic encounters beyond 2-level reconciliation. *Bioinformatics*, https://doi.org/10.1093/bioinformatics/btac062, 2022.

Phylogenomics.fr: a user-friendly web interface to phylogenomic tools and reconciliation workflow

Adrien Limouzin-Lamothe[1], Anne-Muriel Arigon[1] and Vincent Lefort[1,2]

[1] LIRMM UMR 5506, CNRS et Université Montpellier, MONTPELLIER, France
[2] Institut Français de Bioinformatique, CNRS UAR 3601, France

Corresponding Author: adrien.limouzin-lamothe@lirmm.fr

1 Introduction

Phylogenomics aims at reconstructing the evolutionary histories of organisms taking into account whole genomes or large fractions of genomes. Phylogenetic trees reconciliation is a widely used method for reconstructing the evolutionary histories of gene families and species.

Several algorithms were developed toward this goal but the most part of the tool chain is scattered among each developer's repositories. Installing and configuring these heterogeneous programs can be a daunting task for users. Organizing inputs and collecting outputs to run the analyses is usually a time consuming process. To solve this issue, we designed and implemented an online service platform dedicated to phylogenomic analyses.

Our platform provides a user-friendly interface to help biologists to define multiple complex phylogenomic analyses. Its main feature is an automated workflow which takes alignments or gene trees as inputs and which infers reconciliations of gene trees with the species tree. Keeping in mind the concepts that led to the phylogeny.fr (Dereeper A. *et al.*, NAR 2008) and NGphylogeny.fr (Lemoine F. *et al.*, NAR 2019) successes, we designed it as simple and straightforward as possible. We provide automatic workflows dedicated to general users. It can also be fully customized by advanced users.

2 Workflows overview

The workflows are composed with 4 possible steps : gene trees inference, species tree inference, species tree rooting and trees reconciliation. These steps are performed in 4 different use cases (UC) :

- UC 1 : the user provides alignments and chooses the “supertree approach” to infer the species tree
- UC 2 : the user provides alignments and chooses the “supermatrix approach” to infer the species tree
- UC 3 : the user provides his gene trees to infer the species tree with the “supertree approach”
- UC 4 : the user provides his gene trees and his species tree

The user cases 1 and 3 correspond to the automatic workflows that uses default parameters, while the user cases 2 and 4 correspond to the custom workflows.

Valid inputs for the workflows are either a collection of alignments or gene trees and species tree. The platform will parse the input’s type and choose the workflow accordingly.

Gene trees inference step will be processed using PhyML (Guindon S. *et al.*, Syst. Biol. 2010) or FastME (Lefort V. *et al.*, MBE 2015), depending on the input size or the user wish. The user may skip this step by providing his own gene trees.

The second step is to infer the species tree either from the gene trees using the “supertree approach”, or from the alignments using the “supermatrix approach”. Concerning the “supertree approach”, we chose ASTRAL-PRO (Zhang C. *et al.*, MBE 2020) as default but SuperTriplets (Ranwez V. *et al.*, Bioinformatics 2010) and MRP (Ragan M.A., Mol. Phyl. Evol. 1992) are considered as alternatives.

Species tree rooting currently requires a user action (graphical choice of the branch on which the root has to be placed). Automatic alternatives approaches are considered for user convenience.

Once the species tree rooted, we can proceed to the last step. Each gene tree is reconciled with the species tree using ecceTERA (Jacox E. *et al.*, Bioinformatics 2016) : it infers additional evolutionary events such as gene losses, gene duplications and horizontal gene transfers. The final outputs are the reconciled trees saved in RecPhyloXML format (Duchemin W. *et al.*, Bioinformatics 2018). Every steps inputs and outputs can be downloaded and visualized.

SciGeneX: an unsupervised method to naturally discover cell types or cell states based on patterns of co-expressed genes in single-cell RNA-sequencing data

Julie BAVAIS[12], Lionel SPINELLI[12] and Denis PUTHIER[1]

[1] Theories and Approaches of Genomic Complexity (TAGC), Aix-Marseille University, INSERM, Turing Centre for Living Systems, 163 avenue de Luminy, 13009, Marseille, France

[2] Centre d'Immunologie de Marseille Luminy (CIML), Aix-Marseille University, INSERM, CNRS, Turing Centre for Living Systems, 163 avenue de Luminy, 13009, Marseille, France

Corresponding Author: bavais@ciml.univ-mrs.fr

Single-cell RNA sequencing revolutionizes transcriptomic studies providing gene expression level at single-cell resolution [1]. The classical pipeline used to discover cell populations consists of several consecutive steps, namely feature selection, dimension reduction and cell clustering [2]. These steps are widely used in the world of single-cell RNA-sequencing, however, three major problems remain to be improved. First, feature selection methods do not lead to a common consensus and most of them provide variable results [3]. Second, the identification of marker genes leads to double dipping by creating a high type I error rate caused by the prior identification of cell groups [4]. Finally, all the steps of the classical pipeline depend on a number of parameters which, depending on their values, will generate a large variability of results.

To overcome these problems, we developed SciGeneX (for Single-cell informative Gene eXplorer). An unsupervised method offering an alternative approach that provides an initial insight into the pattern of co-expressed genes across cells. SciGeneX automatically filters co-expressed genes across the set of cells using a density-based-filtering algorithm and clusters them into gene patterns of expression using the Markov Cluster Algorithm [5]. Combinations of these patterns spontaneously highlight biologically relevant cell populations associated with cell types or states as well as the genes specifically expressed in these populations. Thus, SciGeneX perform feature selection and identification of co-expressed gene patterns and provide an alternative approach for cell clustering based on these patterns, avoiding the main drawbacks of the currently used algorithms.

Acknowledgements

SciGeneX is freely available as a R package on github at https://github.com/dputhier/scigenex for Linux users and is compatible with the classical seurat workflow.

References

[1] Kolodziejczyk, A.A., Kim, J.K., Svensson, V., Marioni, J.C., and Teichmann, S.A., The technology and biology of single-cell RNA sequencing. Mol. Cell 58, 610–620, 2015.

[2] Luecken MD and Theis FJ. Current best practices in single-cell RNA-seq analysis: a tutorial. Mol Syst iol. 2019 Jun 19;15(6):e8746.

[3] Lähnemann, D., Köster, J., Szczurek, E. et al. Eleven grand challenges in single-cell data science. Genome Biol 21, 31, 2020.

[4] Gao, L. L., Bien, J., and Witten, D., Selective Inference for Hierarchical Clustering, ArXiv, 2020.

[5] Stijn van Dongen, Graph Clustering by Flow Simulation. PhD thesis, University of Utrecht, 2000.

FAIR-Checker: Checking and Inspecting metadata for FAIR bioinformatics resources.

Thomas Rosnet[1,5], Vincent Lefort[2,5], Frédéric De Lamotte[6], Marie-Dominique Devignes[3,5] and Alban Gaignard[4,5]

[1] TAGC/INSERM U1090, Univ Aix-Marseille, Marseille, France
[2] LIRMM, Univ Montpellier, CNRS, Montpellier, France
[3] LORIA, Université de Lorraine, CNRS, Inria, Nancy, France
[4] L'institut du thorax, INSERM, CNRS, University of Nantes, Nantes, France
[5] Institut Français de Bioinformatique, CNRS UAR 3601, France
[6] UMR AGAP Institut, Univ Montpellier, CIRAD, INRAE, Institut Agro, F-34398 Montpellier, France

Corresponding author: thomas.rosnet@france-bioinformatique.fr

1 Introduction

The continuous increase of life science data production raises the importance of better sharing and reusing biological digital resources (datasets, bioinformatics tools or workflows, training materials, etc.). To that end, FAIR principles [1] have been proposed and are being adopted by large communities. However, assessing how much a resource is FAIR is nowadays challenging since answering human-oriented questionnaires is time-consuming and computational evaluations (FAIRMetrics, RDA Maturity Indicators) often require technical expertise.

2 Approach and results

We propose an update of FAIR-Checker [1], aimed at making producers and developers of scientific digital resources more efficient in their FAIR implementation. This tool aims at assessing FAIR principles and providing technical recommendations to enhance the quality of the metadata found the web resources. FAIR-Checker leverages Semantic Web standards and technologies, such as RDF, SPARQL, SHACL, to help users in annotating their resources with high-quality metadata and to ensure interoperability at web scale. It has two main facets: the "Check" module, targeting any user, allows them to execute a list of tests and get a synthetic estimation of the FAIRness of their resource, as well as technical recommendations to improve it. The "Inspect" module, targeting metadata experts, allows them to i) explore and verify if it conforms to community-defined standards and ii) identify missing or non-standard metadata to improve metadata quality.

For each submitted resource, we build a knowledge graph based on embedded RDF triples (microdata, json-ld) as well as external knowledge (public SPARQL endpoints). To evaluate metadata in "Check", we use technologies such as SPARQL queries to automatically assess FAIR metrics, and provide key recommendations in case of failed test with some references to the FAIR-Cookbook [2]. Then, to evaluate metadata quality in "Inspect", we check that used ontology terms are already known in reference registries such as Linked Open Vocabularies (LOV) Ontology Lookup Service (OLS) or BioPortal. Finally, we leverage Bioschemas, the extension of Schema.org for life sciences, to automatically generate SCHACL constraints. Their evaluation informs users on missing metadata, required or recommended for specific types of resources (genes, proteins, training material, computational tools, etc.).

3 Future work

A publication of this work is currently under review. As future work, we aim (i) to support content-negotiation for web sites not embedding metadata in their web pages, (ii) to enhance the matching between a resource type and the corresponding Bioschemas profile, and (iii) to allow the user to annotate and complete its missing metadata. In addition, we plan to develop an API for a better scalability and interoperability of FAIR-Checker.

References

[1] Wilkinson, M., Dumontier, M., Aalbersberg, I. *et al.* The FAIR Guiding Principles for scientific data management and stewardship. Sci Data 3, 160018 (2016).

1. https://fair-checker.france-bioinformatique.fr
2. https://faircookbook.elixir-europe.org/content/home.html

Openlink, a data management dashboard for research teams

Laurent Bouri[1], Mateo Hiriart[2], Guillaume Seith[1], Bertrand Vernay[1], Anne-Cecile Reymann[1], Juliette Godin[1], Nadia Goué[2] and Julien Seiler[1]

[1] IGBMC, 1 Rue Laurent Fries, 67400 Illkirch-Graffenstaden, France
[2] AuBi Platform, Clermont Auvergne Mesocenter, UCA, 63178, Aubière, France

Corresponding Author: laurent.bouri@igbmc.fr

1 Introduction

In the Life Sciences landscape, bioinformatics core facilities play a key role for many scientific communities, by providing software access and reference data in a computational environment tailored for high-throughput computing. They have to handle huge amounts of data generated by scientists, which require ever-increasing resources.

Bioinformatics platforms play a pivotal role to transform scientific raw data into quantitative and integrated analysis, essential to draw conclusions and lead to publications before being made available to the community by deposition in international data warehouses.

In order to help scientists adopt best practices in data management[1], OpenLink[2] has been selected as a "ANR Flash Données Ouvertes" project. It started early 2020 at "Institut de génétique et de biologie moléculaire et cellulaire" (IGBMC) and was renewed by "Institut Français de Bioinformatique" (IFB) in 2021. It involves IGBMC IT department, imaging center, three research teams and IFB collaborators. Openlink's aim is the creation of a web application enabling the establishment of a virtual link between data and metadata scattered over multiple management tools, and thus facilitate the adoption of best practice in data management.

2 OpenLink, an interoperable network of data management tools

Openlink facilitates the transversal identification and manipulation of data associated with a research project and is structured using the ISA model[3] (Investigation, Study, Assay). Openlink is able to connect and create links to any data source and publish them on a repository. The only condition is to have a "Connector" specified in your Openlink project, which dictates how OpenLink communicates with a given data source or repository. OpenLink allows several people to collaborate and share common data without duplication.

From the Data Management Plan (DMP) to data publication, Openlink supports an evolving collection of data sources, including LabGuru, an electronic lab notebook, OMERO, an image visualization, management and data processing tool, Galaxy, a workflow manager, and Zenodo, a universal repository for research outcomes. The goal is to streamline the transfer of data from production to archiving, while automatically enriching data.

3 Conclusion

OpenLink offers dashboards and automatic procedures to support researchers in data management and guide them towards the adoption of a FAIR[4] principle by becoming part of the Open Science movement. Its development is as modular as possible to fit various universities and organization infrastructures and needs.

References

[1] Simms, S., Jones, S., Mietchen, D., & Miksa, T. (2017). Machine-actionable data management plans (maDMPs). Research Ideas and Outcomes, 3, e13086. https://doi.org/10.3897/rio.3.e1308
[2] Openlink source code. https://gitlab.com/igbmc/openlink/openlink
[3] ISA Model and Serialization Specifications. https://isa-specs.readthedocs.io/en/latest/isamodel.html
[4] Directorate-General for Research and Innovation (2018). Turning FAIR into reality. EU publications. https://doi.org/10.2777/1524

IMPatienT: an integrated web application to digitize, process and explore multimodal patient data

Corentin Meyer[1], Norma Romero[2], Teresinha Evangelista[2], Brunot Cadot[3], Jocelyn Laporte[4], Anne Jeannin-Girardon[1], Pierre Collet[1], Kirsley Chennen[1] and Olivier Poch[1]

[1] Complex Systems and Translational Bioinformatics (CSTB), ICube Laboratory, UMR 7357, University of Strasbourg, 1 rue Eugène Boeckel, 67000 Strasbourg, France.

[2] Neuromuscular Morphology Unit, Myology Institute, Reference Center of Neuromuscular Diseases Nord-Est-IDF, GHU Pitié-Salpêtrière, Paris, France

[3] Sorbonne Université, INSERM, Center for Research in Myology, Myology Institute, GHU Pitié-Salpêtrière, Paris, France

[4] Department Translational Medicine, IGBMC, CNRS UMR 7104, 1 rue Laurent Fries, 67404 Illkirch, France.

Corresponding Author: corentin.meyer@etu.unistra.fr

Patient data now incorporates the results of numerous modalities, including imaging, next-generation sequencing and more recently wearable devices. Most of the time, medical acts produce imaging data, such as echography, radiology or histology result in the production of medical reports that describe the relevant findings. Thus, multimodality is induced in patient data, as imaging data is inherently linked to free-text reports.

Useful tools to centralize, process and explore multimodal data are essential to drive research and improve diagnosis. Exploiting patients' data is challenging as the ecosystem of tools is heavily fragmented, depending on the type of data (images, text, genetic sequences), the task to be performed (digitization, processing, exploration) and the domain of interest (clinical phenotype, histology…). To address these challenges, the analysis tools need to be integrated in a simple, comprehensive, and flexible platform.

Here, we present IMPatienT [1] (**I**ntegrated digital **M**ultimodal **PATIEN**t da**T**a), a free and open-source web application to digitize, process and explore multimodal patient data. IMPatienT has a modular architecture, including four components to: (i) create a standard vocabulary for a domain, (ii) digitize and process free-text data by mapping it to a set of standard terms and well-established ontologies, (iii) annotate images and perform image segmentation, and (iv) generate an automatic visualization dashboard to provide insight on the data and perform automatic diagnosis suggestions.

We showcased IMPatienT on a corpus of 89 muscle biopsy reports of congenital myopathy patients provided by the Institute of Myology of Paris and a hundred images of histological sections from the Muscle Atlas. We used the web application to: create a first draft of the muscle histology ontology, digitize the medical reports, and annotate the biopsy images. Exploratory graphs and automatic diagnosis suggestions for the three recurrent classes of congenital myopathies were then automatically generated.

IMPatienT is a web application to digitize, process and explore patient data that can handle image and free-text data. As it relies on user-designed standard vocabulary and well-established ontologies, it is highly flexible to fit any domain of research and can be used as a patient registry for exploratory data analysis (EDA). A demo instance of the application is available at https://impatient.lbgi.fr.

References

[1] C. Meyer et al., "IMPatienT: an integrated web application to digitize, process and explore multimodal patient data." bioRxiv, p. 2022.04.08.487635, Apr. 10, 2022. doi: 10.1101/2022.04.08.487635.

FAIDARE, Plant research data discovery portal for distributed data repositories

Maud Marty[1,2], Célia Michotey[1,2] , Raphaël Flores[1,2] , Jérémy Destin[1,2], Anne Françoise Adam Blondon[1,2] and Cyril Pommier[1,2]

[1] Université Paris-Saclay, INRAE, URGI, 78026, Versailles, France.

[2] Université Paris-Saclay, INRAE, BioinfOmics, Plant bioinformatics facility, 78026, Versailles, France

Corresponding Author: maud.marty@inrae.fr

Plant research, particularly in genetics, often involves identifying heterogeneous and dispersed datasets. Indeed, the study of the behaviour of plants in their environment implies having integrated phenotyping experiment data and cross-referencing it with genetic or molecular variability data (metabolomic, expression, etc.). These data are structured in a very heterogeneous way, according to the specific standards of their field; for exemple MIAME for expression data, MIAPPE [1] for phenotyping data, VCF for genetic variability. Moreover, these data are published in generalist warehouses (Zenodo, data.inrae.fr) or in specialized warehouses (EMBL-EBI) that offer diversified APIs and research interfaces. All this makes the discovery of integrable and interoperable datasets complex.

To simplify this work and maximize the visibility of research data, and their reusability according to FAIR principles, several portal solutions have been developed in recent years, such as omicsDI [2] of the WheatIS (www.wheatis.org). FAIDARE (FAIR Data-finder for Agronomic Research https://urgi.versailles.inrae.fr/faidare/) is a portal dedicated to plant phenomic, genetic and genomic data, allowing to index and standardize heterogeneous metadata sources based on international standards.

FAIDARE offers a simple interface (code in Java, Angular and TimeLeaf) based on a full text search running on Lucene (Elasticsearch). A set of facets allows to refine the search according to the most common criteria: species, data type, data source, ontological annotation, plant material, traits studied, etc. This approach allows the user to easily identify the datasets of interest, to access their description, and then to download them from their source thanks to an export system in csv format. The data is also accessible through a series of REST web services following the Breeding API standard [3].

In this demonstration, we will show how to identify the data needed to study different questions such as the impact of disease on wheat yield or the links between phenology and genetic data in grapevines. We will also see how this network of heterogeneous sources could be harvested, either in a generalist way or by following the Breeding API, via a dedicated Extract Transform Load (ETL) tool coded in python and orchestrated with Nextflow.

Acknowledgements

The web interface has been developed by the Ninja Squad Software development SME (https://ninja-squad.fr/) and URGI. This work was supported by the European Union's Horizon 2020 programme through AGENT (grant agreement No 862613) and GenRes Bridge (grant agreement No 817580) projects, ELIXIR-EXCELERATE (Grant Agreement no. 676559) and Phenome-EMPHASIS (ANR 11-INBS-0012).

References

[1] Evangelia A. Papoutsoglou, Daniel Faria, Daniel Arend, Elizabeth Arnaud, et al. Enabling reusability of plant phenomic datasets with MIAPPE 1.1. New Phytologist (2020). doi: https://doi.org/10.1111/nph.16544

[2] Perez-Riverol Y, et al. Discovering and linking public omics data sets using the Omics Discovery Index. Nat Biotechnol. 2017 May 9;35(5):406-409. doi: https://doi.org/10.1038/nbt.3790.

[3] Peter Selby et al. BrAPI - an application programming interface for plant breeding applications, Bioinformatics (2019). doi: https://doi.org/10.1093/bioinformatics/btz190

Comptage massif et distribué de k-mers

Clément AGRET[1,2], Annie CHATEAU[2] et Alban MANCHERON[2]
[1] Université de Montpellier, LIRMM, France
[2] CIRAD, Montpellier, France

Auteur référent: alban.mancheron@lirmm.fr

https://gite.lirmm.fr/doccy/libGkArrays-MPI

1 Le comptage de k-mers

Les progrès des 15 dernières années en matière de séquençage d'ADN et d'ARN, associés à la baisse de leurs coûts ont eu comme effet direct une production massive de données à analyser. Ce changement d'échelle du volume de données a induit l'apparition de nouvelles méthodologies et notamment celles basées sur le comptage des k-mers – fragments de longueur k – présents dans les séquences. Ces comptages peuvent être utilisés de différentes manières. Par exemple, rechercher des marqueurs spécifiques de certaines populations ou bien pour discriminer les erreurs de séquençage des variations biologiques...

Bien que le comptage de k-mers consiste à associer à une séquence de k nucléotides une valeur entière, il existe de multiples manières de structurer cette information et de l'interroger [1] et les choix algorithmiques et méthodologiques ont une incidence forte sur les performances et la fiabilité des méthodes. De nombreux outils ont été développées pour effectuer ces comptages, tels que `Jellyfish` [2], `DSK` [3], `KMC3` [4], ... Cependant, l'inconvénient commun à la plupart des méthodes existantes actuellement est qu'elles sont limitées par les capacités matérielles de la machine sur laquelle elles sont exécutées. Aussi pour compter les k-mers sur de très gros volumes de données, est-il nécessaire de disposer de machines surpuissantes ou d'adapter les méthodes existantes afin de distribuer les calculs (stratégies *MapReduce* [5]). Nous avons développé une méthode originale permettant de distribuer le calcul sur plusieurs machines, repoussant *de facto* les limitations de ces autres outils.

2 Comptage et indexation massivement parallélisés de k-mers

Nous avons développé une librairie en `C++` (intitulée `libGkArrays-MPI` et distribuée sous la licence libre `CeCILL-C`), exploitant le parallélisme léger (*multithreading*) mais également le calcul distribué, permettant de compter les k-mers des séquences décrites dans un ou plusieurs fichiers (`fasta`, `fastq`, compressés ou non). Outre le simple comptage, cette librairie permet également de les indexer (donc de pouvoir retrouver leurs séquences d'origine). Sur la base de cette librairie, nous avons également développé un outil (intitulé `gkampi` et distribué sous licence libre `CeCILL`) pouvant s'exécuter sur une simple machine comme sur un *cluster* de calcul.

L'outil `gkampi` et la librairie `libGkArrays-MPI` permettent également de compter/indexer des k-mers espacés [6], proposent les même fonctionnalités que les outils standards (`Jellyfish`, `KMC`, ...) et sont documentés. Leur installation est conforme aux standards des `GNU autotools` et le code respecte strictement la norme ISO 2011 du `C++`.

Remerciements

Ce travail a été en partie financé par l'Institut de Biologie Computationnelle de Montpellier, le projet *GenomeHarvest* (*Agropolis foundation*) et le projet *Coalab* (région Languedoc-Roussillon). Nous tenons également à remercier les rapporteurs pour leurs commentaires.

References

[1] Hilde Vinje, Kristian Hovde Liland, Trygve Almøy, and Lars Snipen. Comparing K-mer based methods for improved classification of 16S sequences. BMC *Bioinformatics*, 16:205, 2015.

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

[3] Guillaume Rizk, Dominique Lavenier, and Rayan Chikhi. Dsk: k-mer counting with very low memory usage. *Bioinformatics*, 29(5):652–653, 2013.

[4] Marek Kokot, Maciej Długosz, and Sebastian Deorowicz. Kmc 3: counting and manipulating k-mer statistics. *Bioinformatics*, 33(17):2759–2761, 2017.

[5] Tao Gao, Yanfei Guo, Yanjie Wei, Bingqiang Wang, Yutong Lu, Pietro Cicotti, Pavan Balaji, and Michela Taufer. Bloomfish: A Highly Scalable Distributed k-mer Counting Framework. In *2017 IEEE 23rd International Conference on Parallel and Distributed Systems (ICPADS)*, pages 170–179, December 2017.

[6] Karel Břinda, Maciej Sykulski, and Gregory Kucherov. Spaced seeds improve k-mer-based metagenomic classification. *Bioinformatics*, 31(22):3584–3592, November 2015.

De Novo SNP, InDels and Alternative Splicing Events discovery, annotation and quantification from RNA-seq data with KisSplice

Audric COLOGNE[1], Leandro LIMA[2], Clara BENOIT-PILVEN[3], François GINDRAUD[1], Pierre PETERLONGO[4], Arnaud MARY[1] and Vincent LACROIX[1]

[1] UMR CNRS 5558 BBE, INRIA Erable, Lyon 1 University, 69622, Villeurbanne, France
[2] EMBL-EBI, Welcome Trust Sanger Institute, CB10 1SD, Hinxton, United-Kingdom
[3] FIMM, Helsinki University, 00014, Helsinki, Finland
[4] Univ Rennes, Inria Genscale, CNRS, IRISA—UMR 6074, Rennes, France

Corresponding Author: vincent.lacroix@univ-lyon1.fr

Most eukaryotic genes are composed of exons and introns. At the splicing step, introns are removed from the pre-mRNA and exons are joined together to form a mature mRNA. The precise recognition of exons and introns by the splicing machinery is a complex process, which can be regulated, leading to the production, from a single gene, of possibly several mRNAs with vary in their exon composition, a process known as alternative splicing (AS). The interest in AS has been growing in recent years as more and more studies found this process to be widespread. Deregulation of splicing was also found to be associated to various pathologies.

Transcriptome sequencing can be used to identify and quantify hopefully all spliceforms of all the genes expressed in a particular condition. Comparing biological conditions (patients Vs controls, tissue 1 Vs tissue 2) then enables to identify which genes are differentially spliced. The bioinformatics analysis of RNAseq data remains however challenging in particular when no (good) reference genome is available, or when it is only partially annotated.

Since 2012, we are developping KisSplice (kissplice.prabi.fr), a local de novo assembler which takes as input a list of RNAseq Illumina fastq files and outputs a list of variations (SNPs, indels, AS events) seen in the data. Each variant is quantified in each input sample. Since the first version, we have improved the performance of KisSplice and we developed modules to facilitate the exploration of the results. KissDE is a bioconductor package enabling to assess if a variant is significantly enriched in a condition. KisSplice2RefGenome enables to locate and annotate the variation on a user-provided reference genome. The originality of our method is that its first step is not to map the reads to a reference genome. Instead, it locally assembles the reads, i.e. builds a de Bruijn Graph and searches for specific patterns called bubbles in this graph. We showed that this facilitates the discovery of non-annotated splicing events even when an annotated reference genome is available [2].

In this demo, we will present how to use *KisSplice* in the context of two case studies: (1) the Taybi-Linder Syndrom (TALS), a rare multi-developmental human pathology caused by mutations in one of the minor spliceosome component [5], where *KisSplice* identifies genes that are mis-spliced in patients; and (2) a non-model species, the bug *Rhodnius prolixus* which is an important vector of the Chagas disease, where *KisSplice* identifies genes whose splicing is significantly affected after feeding.

The software is now available through a docker image: https://hub.docker.com/r/dwishsan/kissplice-pipeline, and comes with a Shiny interface which facilitates the exploration of the results.

KisSplice runs within a few hours for datasets up to 1G reads. It requires 30Go of RAM.

References

[1] Hélène Lopez-Maestre *et al.* SNP calling from RNA-seq data without a reference genome: identification, quantification, differential analysis and impact on the protein sequence. *Nucleic Acids Research*, Volume 44, Issue 19, 2 November 2016.
[2] Clara Benoit-Pilven *et al.* Complementarity of assembly-first and mapping-first approaches for alternative splicing annotation and differential analysis from RNAseq data. *Sci Rep* 8, 4307 (2018).
[3] Usama Ashraf, *et al.* Influenza virus infection induces widespread alterations of host cell splicing, NAR Genomics and Bioinformatics, Volume 2, Issue 4, December 2020.
[4] Gustavo Sacomoto *et al.* KISSPLICE: de-novo calling alternative splicing events from RNA-seq data. *BMC Bioinformatics* 13, S5 (2012).
[5] Audric Cologne *et al.* New insights into minor splicing—a transcriptomic analysis of cells derived from TALS patients. *Rna* 25.9 (2019).

RepetDB v2: A unified resource for transposable element references

Mariène WAN[1,2], Nicolas FRANCILLONNE[1,2], Raphaël FLORES[1,2], Françoise ALFAMA[1,2], Johann CONFAIS[1,2], Joëlle AMSELEM[1,2] and Nathalie CHOISNE[1,2]

[1] URGI INRAE Université Paris-Saclay, RD10 Route de Saint-Cyr, 78026, Versailles, France
[2] Université Paris-Saclay, INRAE, BioinfOmics, Plant bioinformatics facility, 78026, Versailles, France

Corresponding Author: nathalie.choisne@inrae.fr

Transposable elements (TEs) are major players in the structure and evolution of eukaryote genomes. Thanks to their ability to move around and replicate within genomes, they are probably the most important contributors to genome plasticity. The insertion of TEs close to genes can affect gene structure, expression and function, contributing to the genetic diversity underlying species adaptation. Many studies have shown that TEs are generally silenced through epigenetic defense mechanisms, and that these elements play an important role in epigenetic genome regulation. Their detection and annotation are considered essential and must be undertaken in the frame of any genome sequencing project.

Here, we will present the new version of RepetDB [1] (Amselem et al., Mobile DNA, 2019), (https://urgi.versailles.inrae.fr/repetdb) our TE database developed to store and retrieve detected, classified and annotated TEs in a standardized manner. This RepetDB v2 new version was updated with 31 more species of plants and fungi and provides TE consensi with evidences able to justify their classification.

RepetDB v2 is a customized implementation of InterMine [2,3], an open-source data warehouse framework used here to store, search, browse, analyze and compare all the data recorded for each TE reference sequence. InterMine provides powerful capabilities to query and visualize all biological information on TE. It allows to make simple search on the database using the QuickSearch (‘google like search’) or make more complex queries using the Querybuilder to display various desired information.

RepetDB v2 is designed to be a TE knowledge base populated with full de novo TE annotations of complete (or near-complete) genome sequences. Indeed, the description and classification of TEs facilitates the exploration of specific TE families, superfamilies or orders across a large range of species. It also makes possible cross-species searches and comparisons of TE family content between genomes.

References

1. Amselem, J., Cornut, G., Choisne, N., Alaux, M., Alfama-Depauw, F., Jamilloux, V., Maumus, F., Letellier, T., Luyten, I., Pommier, C., Adam-Blondon, A. F., & Quesneville, H. (2019). RepetDB: a unified resource for transposable element references. Mobile DNA, 10, 6. https://doi.org/10.1186/s13100-019-0150-y
2. InterMine: extensive web services for modern biology. Kalderimis A, Lyne R, Butano D, Contrino S, Lyne M, Heimbach J, Hu F, Smith R, Stˇepán R, Sullivan J, Micklem G. Nucleic Acids Res. 2014 Jul; 42 (Web Server issue): W468-72
3. InterMine: a flexible data warehouse system for the integration and analysis of heterogeneous biological data. Smith RN, Aleksic J, Butano D, Carr A, Contrino S, Hu F, Lyne M, Lyne R, Kalderimis A, Rutherford K, Stepan R, Sullivan J, Wakeling M, Watkins X, Micklem G. Bioinformatics (2012) 28 (23): 3163-3165.

[Jeudi 7 - 12h00 - Salle 15 ; Thursday 7th - 12 :00 - room 15]

Snakemake RNAseq workflow including repeat expression analysis

Ali Hamraoui and Magali Hennion
Epigenetics and Cell Fate, Université Paris Cité, CNRS, 35 rue Hélène Brion, 75013, Paris, France

Corresponding author: magali.hennion@cnrs.fr

1 Introduction

Repeated elements constitute a huge fraction of many genomes, with about half of the sequence consisting of repetitive elements in humans [1]. A significant part of those sequences are transcribed and transcription misregulation is associated with several diseases, including rare genetic diseases and cancers. Most RNAseq analyses ignore repeated elements, but several tools have been developed to specifically look at their transcription [2]. Here we implemented two approaches, one based on featureCounts [3] and one on TEtranscripts [4], into a complete analysis workflow coded with Snakemake [5] and based on RASflow [6]. The workflow is usable by biologists with no bioinformatics background, and is also routinely used for standard gene expression analysis.

2 Workflow description

The workflow is written in Python and uses Snakemake in a dedicated Conda environment. It was optimized to compute efficiently the data on HPC clusters using Slurm such as IFB-core or iPOP-UP clusters. It includes the following steps:

- Download FASTQ files from SRA (facultative)
- Quality control of the raw data
- Trimming of low quality reads, adapters and/or of a specific number of bases (facultative)
- Mapping and QC
- Counting and QC
- Differential expression analysis (genes and/or repeats)
- Html report with interactive plots

Different tools were implemented for each step and the user can configure the workflow thanks to a simple yaml file.

Documentation: *https : //parisepigenetics.github.io/bibs/edctools/workflows/rasflow_edc*

Acknowledgements

We thank Jean-François Ouimette and Guillaume Velasco for the co-supervision of Ali's work, constructive discussions and data. We thank Madeleine Moscatelli for providing the data used to test the pipeline. We acknowledge Olivier Kirsh and all BiBs steering committee for helpful discussions, and Julien Rey and all iPOP-UP technical committee for setting up the cluster.

References

[1] S. Nurk, S. Koren, A. Rhie, M. Rautiainen, A. V. Bzikadze, A. Mikheenko, M. R. Vollger, and colleagues. The complete sequence of a human genome. *Science*, 376(6588):44–53, 04 2022.

[2] S. Lanciano and G. Cristofari. Measuring and interpreting transposable element expression. *Nat Rev Genet*, 21(12):721–736, 12 2020.

[3] Y. Liao, G. K. Smyth, and W. Shi. featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. *Bioinformatics*, 30(7):923–930, Apr 2014.

[4] Y. Jin, O. H. Tam, E. Paniagua, and M. Hammell. TEtranscripts: a package for including transposable elements in differential expression analysis of RNA-seq datasets. *Bioinformatics*, 31(22):3593–3599, Nov 2015.

[5] F. Mölder, K. P. Jablonski, B. Letcher, M. B. Hall, C. H. Tomkins-Tinch, V. Sochat, J. Forster, S. Lee, S. O. Twardziok, A. Kanitz, A. Wilm, M. Holtgrewe, S. Rahmann, S. Nahnsen, and J. Köster. Sustainable data analysis with Snakemake. *F1000Res*, 10:33, 2021.

[6] X. Zhang and I. Jonassen. RASflow: an RNA-Seq analysis workflow with Snakemake. *BMC Bioinformatics*, 21(1):110, Mar 2020.

ASTERICS: A Tool for the ExploRation and Integration of omiCS data.

Élise Maigné[1,*], Céline Noirot[1,2,*], Jérôme Mariette[1,2], Yaa Adu Kesewaah[1,3], Sébastien Déjean[3,4], Camille Guilmineau[1,3], Julien Henry[1,3], Arielle Krebs[1,2], Laurence Liaubet[5], Fanny Mathevet[1,3], Hyphen-Stat[6], Christine Gaspin[1,3], and Nathalie Vialaneix[1]

[1] Université de Toulouse, INRAE, UR MIAT, 31326, Castanet-Tolosan, France
[2] Université Fédérale de Toulouse, INRAE, BioinfOmics, GenoToul Bioinformatics facility, 31326, Castanet-Tolosan, France
[3] Plateforme Biostatistique, Genotoul, Toulouse, France
[4] IMT, UMR5219, Université de Toulouse, CNRS, UPS, 31062, Toulouse, France
[5] GenPhySE, Université de Toulouse, INRAE, ENVT, F-31326, Castanet-Tolosan, France
[6] Hyphen-stat, https://hyphen-stat.com/, Toulouse, France
[(*)] These authors contributed equally to the work.

Corresponding author: nathalie.vialaneix@inrae.fr

1 Introduction

The rapid development of omics acquisition techniques has induced the production of a large volume of heterogeneous and multi-level omics datasets measured on the same individuals. Complex information of biological interest is obtained from so-called *integration methods*, which have been increasingly developed in the past few years. Some of these methods are already available in R packages (like **mixOmics** [1] or **mixKernel** [2] to which our team has contributed). However, the use of these packages still requires to learn a programming language and to have access to sufficient statistical knowledge to choose method parameters and interpret outputs.

2 ASTERICS

ASTERICS is a web application that aims at making complex exploratory and integration analysis workflows easily available to biologists. Data edition, exploration and integration menus organize the interface to perform 1/ data edition⋆, missing value imputation, and normalization⋆, 2/ data exploration with interactive plots, numerical summaries, PCA, tests, clustering, and self-organizing maps, and 3/ data integration with differential analysis⋆, MFA, or PLS-based methods. Analyses are adapted⋆ to the most standard omics datasets (RNA-seq or count data from sequencing technologies, microarray, metabolomics, metagenomics or other compositional data).

ASTERICS is also designed to make the analysis flow understandable with a navigable workspace that displays uploaded or obtained datasets and performed analyses in a graph. Finally, it also comes with a documentation for beginners⋆ that helps interpret the results, choose proper options or the next analysis to perform.

ASTERICS is based on Rserve, pyRserve, and flask. R package versions are controlled using **renv**. Frontend is developed in Vue.js and uses the CSS framework Bulma.

A first and limited version of ASTERICS is already available online at http://asterics.miat.inrae.fr/. This limited version does not include the features highlighted above with the mark "⋆" at time of writing of this proposal. ASTERICS will also be released as a docker image. The complete production version of ASTERICS is scheduled for September 2022, with intermediate versions, including an increasing number of features, deployed online meanwhile.

Acknowledgements

This work is funded by Région Occitanie (Grant 20008788 – ASTERICS).

References

[1] Florian Rohart, Benoît Gautier, Amrit Singh, and Kim-Anh Le Cao. mixomics: an R package for 'omics feature selection and multiple data integration. *PLoS Computational Biology*, 13(11):e1005752, 2017.

[2] Jérôme Mariette and Nathalie Villa-Vialaneix. Unsupervised multiple kernel learning for heterogeneous data integration. *Bioinformatics*, 34(6):1009–1015, 2018.

Fantasio: Identifying rare recessive variants involved in multifactorial traits

Margot DEROUIN[1], Sidonie FOULON[1], Marie-Sophie OGLOBLINSKY[2], Steven GAZAL[3], Hervé PERDY[4] and Anne-Louise LEUTENEGGER[1]

1 Inserm, Université Paris Cité, NeuroDiderot, UMR1141, 48 bd Sérurier, 75019, Paris, France

2 Inserm, Université de Bretagne Occidentale, GGB, UMR1078, 22 avenue Camille Desmoulins, 29200, Brest, France

3 University of Southern California, 1450 Biggy Street, 90033, Los Angeles, USA

4 Inserm, Université Paris Saclay, CESP, UMR1018, 16 Avenue Paul-Vaillant-Couturier, 94807, Villejuif, France

Corresponding Author: margot.derouin@inserm.fr

For more than a decade, genome-wide association studies (GWAS) have made it possible to detect associations between genetic variants and complex diseases in population samples. Their experimental design uses mostly variants that are common in the population, and studies them according to an additive genetic model [1,2]. It appears however that the genetic component of multifactorial diseases is not yet fully elucidated, which could be partly due to the contribution of rare variants with recessive effects, not detected by classic GWAS. These types of variants can be found in HBD-segments.

The original HBD-GWAS strategy [3] points out different regions of interest for the association with common complex traits in different populations. It is based on the excess of homozygosity by descent (HBD) segments shared by *consanguineous cases only* as in homozygosity mapping and thus define candidate regions that may contain rare recessive variants.

We propose here an extension to the HBD-GWAS approach to identify rare recessive variants. This approach relies on an excess of homozygous-by-descent segments shared *among cases compared to what is expected among controls.* We have implemented it in an R package, Fantasio. We illustrate its performance on the UK Biobank cohort (~500,000 individuals living in the UK). We focus on the diabetes phenotype constructed from available fields of the biobank.

The HBD-GWAS strategy points out different regions of interest for the association with diabetes in the different populations. We show how the extension of the approach to include the HBD segments from the controls allows to discriminate between these different signals.

Acknowledgements

This research work was conducted using the UK Biobank biomedical database www.ukbiobank.ac.uk under Application #59366 - Method developments for the genetic analysis of complex traits and was funded by the Inserm cross-cutting program GOLD (GenOmics variability in health & Disease).

References

[1] Welter D, MacArthur J, Morales J, Burdett T, Hall P, Junkins H, et al. The NHGRI GWAS Catalog, a curated resource of SNP-trait associations. Nucleic Acids Res. 2014 Jan;42(D1):D1001–6.

[2] Loos RJF. 15 years of genome-wide association studies and no signs of slowing down. Nat Commun. 2020 Nov;11(1):5900.

[3] Génin E, Sahbatou M, Gazal S, Babron M-C, Perdry H, Leutenegger A-L. Could Inbred Cases Identified in GWAS Data Succeed in Detecting Rare Recessive Variants Where Affected Sib-Pairs Have Failed? Hum Hered. 2012;74(3–4):142–52.

Visualization of FAANG data with VizFaDa

Laura Morel[1], Peter Harrison[2] and Guillaume Devailly[1]

[1] GenPhySE, 24 chemin de Borde-Rouge, 31326, Castanet-Tolosan, France
[2] EMBL-EBI, Hinxton, United Kingdom

Corresponding Author: laura.morel@inrae.fr

The FAANG (*Functional Annotation of Animal Genomes*) international consortium aims to produce high-quality functional annotation of the genomes of domesticated animals [1]. Members of the community can submit their epigenomics, transcriptomics or genomics data to the FAANG Data Portal (https://data.faang.org) coordinated by a Data Coordination Centre at the EMBL-EBI [2]. FAANG data conforms to principles of findability, accessibility, interoperability and reusability (FAIR). The FAANG Data Portal allows users to find, select and download datasets relevant to their research using extensive sample and experimental metadata standards.

VizFaDa aims to produce interactive data visualization through web applications intended to be integrated to the FAANG Data Portal. In order to generate those visualizations, the raw data from the portal has to be processed. During this step, quality control reports are created, providing valuable and previously unavailable insight into the quality of the data. VizFaDa focuses on RNA-seq, ChIP-seq and DNA methylation data.

Interactive clustered correlation heatmap are generated, allowing the user to compare experiments from a certain assay type within a species. Experiments with similar results are clustered together. The user can use FAANG metadata to annotate the heatmap or to filter experiments from the database for a more focused visualization. Stacked epigenetic profiles are created from gene expression and epigenetic data obtained either from the same sample or from two comparable samples, notably at transcription start sites. This allow the investigation of relationship between epigenetic marks and transcription levels. Data submitted to the portal will be automatically processed and added to VizFaDa, ensuring the long-term relevance and accuracy of the project.

During VizFaDa demonstration, I will be presenting features and several use-cases of the VizFaDa web application, and discuss how the community can take advantage of our work.

Acknowledgements

The authors would like to thank Alexey Sokolov from EMBL-EBI, Sylvain Foissac and the GenEpi team at GenPhysSE for their support.

References

[1] The FAANG Consortium; Andersson, L.; Archibald, A. L.; Bottema, C. D.; Brauning, R.; Burgess, S. C.; Burt, D. W.; Casas, E.; Cheng, H. H.; Clarke, L.; et al. Coordinated International Action to Accelerate Genome-to-Phenome with FAANG, the Functional Annotation of Animal Genomes Project. *Genome Biology,* (16):57, 2015

[2] Harrison, P. W.; Fan, J.; Richardson, D.; Clarke, L.; Zerbino, D.; Cochrane, G.; Archibald, A. L.; Schmidt, C. J.; Flicek, P. FAANG, Establishing Metadata Standards, Validation and Best Practices for the Farmed and Companion Animal Community. *Animal Genetics*, (49):520–526, 2018

ProFeatMap: a customizable tool for 2D feature representation of protein sets

Goran BICH[1], Elodie MONSELLIER[1], Gilles TRAVE[1], Yves NOMINE[1]

[1] Institut de Génétique et de Biologie Moléculaire et Cellulaire (IGBMC), 1 rue Laurent Fries, 67404, Illkirch, France

Corresponding Author: bichg@igbmc.fr

Application: https://profeatmap.pythonanywhere.com/

Many biological studies, particularly –omics, involve lists of proteins or related macromolecules. Links between a list's components can be found by analyzing their protein-associated biological processes by Gene Ontology terms or by comparing their sequences. While Gene Ontology might be able to find over-represented features in the protein list, it loses information of their relative location, size and organization. Alternatively, sequence examination implies complex analyses such as multiple alignments and requires similar proteins to be informative. An intermediate scale of analysis is to focus on features such as domains, amino-acid or domain repeats, post-translational modifications, variants, secondary structures, low-complexity regions, and their organization along the sequence. The web interface ProFeatMap has been developed for feature visualization of protein datasets, as maps, in a highly customizable way, to investigate proteins on a more global scale.

The main usage of ProFeatMap is the creation of maps, duh. Based on a user-defined list of proteins, information is collected from the Uniprot database. This data is used to find most represented features in the list and to choose a shape and color for each. Most commonly found features (domains and repeats) keep a consistent representation across maps. The protein list is then sorted, clustering protein with similar feature content. These default parameters are used to create maps within minutes and can give several types of insights: General structural and functional organization of features, highlighting of conserved features or feature patterns, of evolutionary link between proteins or potential annotation issues and current state of available structural data such as experimental coverage (PDBs) and secondary structure. This new knowledge and additional investigations can then be displayed by tinkering with various parameters. ProFeatMap gives full freedom for removing and adding features, changing their graphical representation and modify general map parameters, not requiring any prior programming knowledge.

ProFeatMap also include more advanced tools that can be used to investigate previously highlighted points of interest. First of all, after data collection from Uniprot, all extracted data is compiled in a single file containing: A list of all features position and size, the occurrence of each feature, a list of all found PDB structure, the length of each protein and its sequence. The user can easily retrieve all sequences for a feature of interest as a fasta file, compatible with multiple alignments tools. This option can be used to rapidly assess potential annotation issues or to highlight conserved residues in the feature. Additionally, it is possible to search for features, principally motifs, using regular expressions, rapidly returning the list of hits, which can then be added to a map if wished. Finally, it is worth to notice, that ProFeatMap automatically searches for protein names if only Uniprot codes are inputted in the list. This will not only identify obsolete or invalid codes but also find corresponding organisms.

Posters

EstiAge: A tool to estimate the age of a variant

Thomas E. LUDWIG[1,2] and Emmanuelle GÉNIN[2]

[1] CHU Brest, Brest 29200, France

[2] INSERM, Université de Bres, EFS, UMR 108 GGB, Brest 29200, France

Corresponding Author: thomas.ludwig@inserm.fr

1. Introduction

In the last decades, a huge number of genetic variants associated to diseases and phenotypes have been identified. By observing linked genetic markers, it is possible to estimate the number of generations that separate the current carriers from their supposed common ancestor, from whom they inherited this variant. These estimations are particularly relevant to understand disease spread in populations. They confirm that rare variants are often relatively recent.

2. Implementation

EstiAge [1] relies on linkage disequilibrium (LD) decay to estimate the age (in number of generations) of the most recent common ancestor (MRCA) carrying this variant. The idea is to measure the length of the haplotype containing this variant that is shared by all the current carriers. The shorter the haplotype, the older the variant. To do so, a set of markers (microsatellites) shared by all the carriers is used. The allelic differences at the individual level for these markers serve to establish the limits of the common haplotype.

3. Improvements

This new version of EstiAge aims to simplify user experience and is available as a webservice on https://lysine.univ-brest.fr/estiage where the user can provide microsatellite data for a set of samples carrying the variant of interest. The service will then construct an EstiAge input file and subsequently estimate the variant's age. EstiAge can also be downloaded as a Java Archive so as to be executed locally in order to preserve data privacy.

As pointed by [2], the main drawback of EstiAge was its restriction to microsatellite data. In this version, we addressed this issue and EstiAge can now also be used on the SNPs/INDELs contained in a VCF file. In this case, as proposed in [3], homozygous variants will serve as markers during the haplotypes reconstruction.

4. Use Case

EstiAge was successfully used to estimate the age of the Phe508del mutation of the CFTR gene and involved in Cystic Fibrosis [4].

Acknowledgements

We thank Thomas Martiny and Tom Guyader for their contribution to this work.

References

1. E. Genin, A. Tullio-Pelet, F. Begeot, S. Lyonnet, et L. Abel. Estimating the age of rare disease mutations: the example of Triple-A syndrome. *J Med Genet*, vol. 41, nº 6, p. 445-449, juin 2004, doi: 10.1136/jmg.2003.017962.
2. L. C. Gandolfo, M. Bahlo, et T. P. Speed. Dating rare mutations from small samples with dense marker data. *Genetics*, vol. 197, nº 4, p. 1315-1327, août 2014, doi: 10.1534/genetics.114.164616.
3. I. Mathieson et G. McVean. Demography and the Age of Rare Variants. *PLoS Genet*, vol. 10, nº 8, p. e1004528, août 2014, doi: 10.1371/journal.pgen.1004528.
4. P. Farrell *et al.* Estimating the age of p.(Phe508del) with family studies of geographically distinct European populations and the early spread of cystic fibrosis. *Eur J Hum Genet*, vol. 26, nº 12, p. 1832-1839, déc. 2018, doi: 10.1038/s41431-018-0234-z.

ROCK: digital normalization of whole genome sequencing data

Véronique LEGRAND[1], Thomas KERGROHEN[2,3], Nicolas JOLY[1] and Alexis CRISCUOLO[4,5]

[1] Institut Pasteur, Université Paris Cité, Plateforme HPC, F-75015 Paris, France
[2] Prédicteurs moléculaires et nouvelles cibles en oncologie, INSERM, Gustave Roussy, Université Paris-Saclay, Villejuif, France
[3] Département de Cancérologie de l'Enfant et de l'Adolescent, Gustave Roussy, Université Paris-Saclay, Villejuif, France
[4] Institut Pasteur, Université Paris Cité, Bioinformatics and Biostatistics Hub, F-75015 Paris, France
[5] Institut Pasteur, Université Paris Cité, Plateforme de Microbiologie Mutualisée (P2M), F-75015 Paris, France

Corresponding Author: alexis.criscuolo@pasteur.fr

Abstract

Due to advances in high-throughput sequencing technologies, generating whole genome sequencing (WGS) data with high coverage depth (e.g. $\geq 500\times$) is now becoming common, especially when dealing with non-eukaryotic genomes. Such high coverage WGS data often fulfills the expectation that most nucleotide positions of the genome are sequenced a sufficient number of times without error. However, performing bioinformatic analyses (e.g. sequencing error correction, whole genome *de novo* assembly) on such highly redundant data requires substantial running times and memory footprint.

To reduce redundancy within a WGS dataset, randomly downsampling high-throughput sequencing reads (HTSR) is trivial. Nevertheless, this first-in-mind strategy is not efficient as it does not minimize variation in sequencing depth, thereby eroding the coverage depth of genome regions that are under-covered (if any). To cope with this problem, a simple greedy algorithm, named *digital normalization*, was designed to efficiently downsample HTSRs over genome regions that are over-covered [1]. Given an upper-bound threshold $\kappa > 1$, it returns a subset of HTSRs inducing an expected coverage depth of at most $\varepsilon\kappa$ across the genome (where $\varepsilon > 1$ is a small constant). By discarding highly redundant HTSRs while retaining sufficient and homogeneous coverage depth ($\approx \varepsilon\kappa$), this algorithm strongly decreases both running times and memory required to subsequently analyze WGS data, with often little impact on the expected results.

Interestingly, the *digital normalization* algorithm can be easily enhanced in several ways, so that the final subset contains fewer but more qualitative HTSRs. ROCK (*Reducing Over-Covering K-mers*) was therefore developed with the key purpose of implementing a fast, accurate and easy-to-use *digital normalization* procedure. Developed in C++, ROCK enables to observe fast running times using only a unique thread. To improve the *digital normalization* procedure, ROCK also implements two novel strategies: (i) downsampling the HTSRs based on their Phred scores, and (ii) implementing a final step that filters out low-covering HTSRs. Thanks to these improvements, ROCK [2] can be used as a preprocessing step prior to performing fast genome *de novo* assembly. The source code is available under GNU Affero General Public License v3.0 at `https://gitlab.pasteur.fr/vlegrand/ROCK`.

Acknowledgements

We acknowledge the help of the HPC Core Facility of the Institut Pasteur for this work.

References

1. C. Titus Brown, Adina Howe, Qingpeng Zhang, Alexis B. Pyrkosz, and Timothy H. Brom. A Reference-Free Algorithm for Computational Normalization of Shotgun Sequencing Data. *arXiv*, 1203.4802v2, 2012.
2. Véronique Legrand, Thomas Kergrohen, Nicolas Joly, and Alexis Criscuolo. ROCK: digital normalization of whole genome sequencing data. *Journal of Open Source Software*, 7(73):3790, 2022.

msscaf: a multiple source genome scaffolder

Matthias Zytnicki
MIAT, INRAE, Castanet-Tolosan, France

Corresponding author: matthias.zytnicki@inrae.fr

1 Introduction

The aim of the assembly is to build a given genome from sequenced reads. New technologies, which include ONT and PacBio, provide long reads which are crucial to longer assemblies. After sequencing, reads are merged into longer fragments, called contigs. These contigs are usually fragmented, and are not sufficient to provide telomere to telomere assemblies. Several methods can then be used in order to scaffold these contigs, which order and orient them.

Among them, linked reads, such are 10X Genomics, and Hi-C are the most widely used. The first type of data make it possible to connect contigs distant by at most 100kbp. Hi-C can make longer joins, but is more fuzzy. Dedicated tools for linked reads and Hi-C are already available. However, each one sometimes make choices that are clearly contradicted by the other type of data. Our aim here is to provide a unified tool, which would simultaneously leverage all types of information, including the long reads themselves.

2 Results

Our method first place each type of information (long reads, linked reads, Hi-C) into bins of fixed size. These bins are then stored into sparse matrices. We then proceed with 3 steps.

Parameter estimation We first possibly merge bins into larger, meta-bins, when matrices are too sparse to be used. We estimate the length of each signal: the molecule size for long reads and linked reads, the maximum expected contact length for Hi-C data. We also compute the decay, which is expected number of contacts given a distance between two genomic positions. We finally compute and discard outlier bins, which contain too much or too few counts.

Splits We detect splits, which are erroneous contig connections. These splits are found while examining the decay function, which is unexpectly low in these positions. The splits found with one type of data are compared with the signal found in other types. If a split is contradicted by other data, it is dropped.

Joins Joins are detected by finding decay-like signal in the corners of the matrices involving the contigs to be joined. For a given contig end, we compare the different possible joins. If one join is clearly much stronger than the other ones, it is used. In case of doubt, the joins are dropped.

3 Conclusion

The tool has been developped in R/C++, and is available in https://github.com/mzytnicki/msscaf. It can generate several figures, so that the user can visually validate the splits/joins suggested by the method.

It is currently being compared to other methods.

Acknowledgements

This work has been supported by the SeqOccIn program https://get.genotoul.fr/seqoccin/, and financed by FEDER funds (Programme Opérationnel FEDER-FSE Midi-Pyrénées et Garonne 2014-2020).

Évaluation comparative des méthodes d'alignement multiple de séquences appliquées au séquençage de troisième génération

Coralie ROHMER[1], Hélène TOUZET[1] et Antoine LIMASSET[1]
Univ. Lille, CNRS, Centrale Lille, UMR 9189 - CRIStAL, 59000 Lille, France

Auteur référent: coralie.rohmer@univ-lille.fr

Le séquençage de troisième génération modifie radicalement la manière d'appréhender l'accès à l'information génomique. La possibilité d'obtenir des reads longs de dizaines ou de centaines de kilobases permet de discerner des structures génomiques complexes et d'appréhenderla structure globale des génomes. Cependant, ces données présentent une grande quantité de bases erronées, y compris des délétions et des insertions [1]. Des nouveaux outils et de nouvelles méthodes ont été spécifiquement développés pour traiter ces erreurs et obtenir des séquences fiables à partir de ces long reads [2,3,4,5,6]. Mais, à ce jour, l'élimination complète des erreurs reste encore un problème ouvert. Dans ce travail, nous examinons si les outils traditionnels d'alignement multiple de séquences (MSA), qui ont été conçus pour traiter des données très différentes, peuvent traiter ces données de séquençage. En d'autres termes, dans quelle mesure les outils de MSA existants peuvent-ils s'adapter au profil d'erreur et à la longueur des long reads ? Pour répondre a cette question, nous avons développé un benchmark qui permet d'évaluer la performance des outils de MSA dans des conditions variables : longueur de la région cible (de 100nt à 10000nt), profondeur de séquençage (de $\times 10$ à $\times 200$), taux d'erreur (de 1% à 30%) et profil d'erreur (proportion de substitutions, insertions et délétions).

Tout d'abord, les lectures sont alignées sur la région cible avec Minimap2 et tronquées pour obtenir des piles de reads couvrant la région. Des MSA sont ensuite construits sur cette sélection de reads avec les différentes outils. Enfin, nous calculons une série de métriques : séquence consensus avec pourcentage de matchs, d'erreurs, d'identités, de caractères ambigus (IUPAC), temps de calcul et ressources mémoire. Ce workflow est développé avec Snakemake. Nous l'avons utilisé pour comparer les outils de MSA les plus populaires avec des stratégies d'alignement complémentaires (POA, AbPOA, SPOA, Muscle, Clustal Omega, T-Coffee, Mafft et Kalign) sur des reads réels issus de séquençages Nanopore (*E.coli* SRR8335315 et SRR12801740, *Saccharomyces cerevisiae* ERR4352154 et ERR4352155, et *Homo sapiens*) ainsi que des reads simulés, pour tester différents profils d'erreurs. Grâce a ce travail, nous avons observé plusieurs comportements intéressants. Tout d'abord, il existe des variations conséquentes en terme de temps et de ressources mémoire entre les différentes méthodes de MSA selon les paramètres. La qualité obtenue est également très différente d'une méthode à l'autre, certaines méthodes étant incapables de produire une précision élevée ou incapables de gérer les cas diploïdes. Toutes ces observations sont nécessaires pour comprendre comment gérer le taux d'erreur des long reads et développer des nouvelles méthodes de correction ou de polissage.

Remerciements

Ce travail est soutenu par la Région Hauts-de-France et l'ANR ASTER (ANR-16-CE23-0001).

Références

[1] R.Krishnakumar et al. Systematic and stochastic influences on the performance of the minion nanopore sequencer across a range of nucleotide bias. *Scientific reports*, (8) :1–13, 2018.

[2] S. Koren et al. Canu : scalable and accurate long-read assembly via adaptive k-mer weighting and repeat separation. *Genome research*, (27) :722–736, 2017.

[3] C.L. Xiao et al. Mecat : fast mapping, error correction, and de novo assembly for single-molecule sequencing reads. *Nature methods*, (14) :1072, 2017.

[4] R. Vaser et al. Fast and accurate de novo genome assembly from long uncorrected reads. *Genome research*, (27) :737–746, 2017.

[5] P. Morisse et al. Consent : Scalable self-correction of long reads with multiple sequence alignment. *BioRxiv*, page 546630, 2019.

[6] R. Warren et al. ntedit : scalable genome sequence polishing. *Bioinformatics*, (35) :4430–4432, 2019.

Preliminary Results from Nanopore Q20+ Sequencing

Jules SABBAN[1], Camille ECHE[1], Joanna LLEDO[1], Amandine SUIN[1], Céline VANDECASTEELE[1], Eden DARNIGE[1], Clement BIRBES[2], Andreea DREAU[2], Christophe KLOPP[2], Christine GASPIN[2], Denis MILAN[1], Carole IAMPIETRO[1], Cécile DONNADIEU[1], Gérald SALIN[1], Céline LOPEZ-ROQUES[1] and Claire KUCHLY[1]

[1] INRAE US 1426 GeT-PlaGe Genotoul, Chemin de Borde Rouge, 31326, Castanet-Tolosan, France

[2] INRAE MIAT UR875 Plateforme Bioinformatique Genotoul, Chemin de Borde Rouge, 31326, Castanet-Tolosan, France

Corresponding Author: claire.kuchly@inrae.fr

Common genomic problems such as complex genome assembly, discovery of structural variants or phasing can be solved using long read technologies.

Implemented since 2015 at the GeT-PlaGe platform, different long read technologies have significantly evolved in this short time, each one having its own strengths and weaknesses. Read quality is the most important parameter for data analysis, and is often considered to be a weak point of long read technologies. To address this issue, Oxford Nanopore Technologies (ONT) has recently developed a new chemistry, referred to as Q20+, that promises long reads with a very low error rate. We are currently testing and evaluating the performance of this new feature at GeT-PlaGe for project SeqOccIn.

We show in detail the library preparation with kit 12, which is used to generate Q20+ reads, and how we have tested it on the GridION. Raw data analysis using the Guppy 6 basecaller demonstrates higher Q-scores for simplex reads as compared to Guppy 5 raw reads, additionally showing a small percentage of duplex reads with Qscores of around Q35. Using an in-house pipeline that combines ONT's MinKnow software, Duplex Tools, the Guppy 6 basecaller, and our own custom scripts, we have generated statistics on alignments produced with minimap2. We found that the newest generation of flow cells (R10.4) and the latest version of Guppy (version 6) improve the alignment quality. However, when comparing the read accuracy of LSK109+R9.4.1 and LSK112+R10.4 using the same software treatment (Guppy 5), no major difference is detected. Oxford Nanopore Technology is currently working to increase the duplex read rate, announcing that we could soon have around 40% duplex reads, which would drastically improve the overall read accuracy.

Acknowledgements

We thank the GenoToul bioinformatics facility for their support in computing resources and data storage, and for helping us process the data and making available all the needed software and infrastructure.

We thank the Genoscope with whom we have always had very fruitful exchanges and who shared their test script with us.

SVJedi-graph: genotyping close and overlapping structural variants with a variation graph and long-reads

Sandra Romain[1] and Claire Lemaitre[1]
Université de Rennes 1, Inria, IRISA, 35000, Rennes, France

Corresponding author: claire.lemaitre@inria.fr

Abstract

Structural variants (SVs) are genomic segments of more than 50 bp that have been rearranged in the genome. The advent of long-read sequencing technologies has increased and enhanced their study, and a great number of SVs has already been discovered in many species. Complementary to their discovery, the genotyping of known SVs in newly sequenced individuals is of particular interest for several applications such as trait association and clinical diagnosis. Due to SVs' large size range (up to a few megabases), long-reads are more suited for their study than short-reads. As such, our team previously released SVJedi [1], one of the first SV genotypers using long-read data. SVJedi's method of representing independently both SV's allelic sequences reduced reference bias in genotyping and showed improved genotyping performances. However, the method failed to genotype closely located or overlapping SVs due to redundancy in representative allelic sequences.

To overcome this limitation, we present SVJedi-graph, a long-read SV genotyper based on a variation graph to represent SV alleles. The use of sequence graphs to represent SVs for genotyping is fairly recent [2,3,4,5], but existing methods are restricted to short-read data, and SVJedi-graph is the first graph-based SV genotyper using long-reads. In our method, we build the variation graph from a reference genome and a given set of SVs. The genome sequence is split in fragments at each SV's start and end positions, and each fragment becomes a node in the graph. Edges are added between nodes to indicate reference and alternative paths for each SV, and additional nodes are added for insertions. Then, the long reads are mapped on the variation graph using GraphAligner [6] and the resulting alignments are filtered on their quality and mapping localization. Finally, the most likely genotype for each SV is predicted from the ratio between the number of reads supporting each allele.

SVJedi-graph can genotype four SV types as of now, namely deletions, insertions, inversions and translocations. Running SVJedi-graph on simulated sets of deletions showed that the use of a variation graph was able to restore the genotyping quality on close and overlapping SVs. For instance, with a simulated set of deletions that had another close deletion 0 to 50 bp apart, we obtained a genotyping rate (proportion of SVs with a predicted genotype) of 99.9% and an accuracy (proportion of accurate genotype predicted among all predicted genotypes) of 99.0%, compared to a genotyping rate of 78.9% and an accuracy of 97.3% with SVJedi on the same dataset. We also tested our method on the real gold standard dataset of Genome In A Bottle (human individual HG002), and were able to obtain a higher genotyping rate than SVJedi on the same data (97.4% against 90.2%), with a similar or slightly better accuracy (92.9% against 92.2%). SVJedi-graph is distributed under an AGPL license and available on GitHub at https://github.com/SandraLouise/SVJedi-graph.

References

[1] L. Lecompte et al. SVJedi: genotyping structural variations with long reads. *Bioinformatics*, 36(17):4568–4575, 2020.

[2] E. Garrison et al. Variation graph toolkit improves read mapping by representing genetic variation in the reference. *Nature Biotechnology*, 36(9):875–879, 2018.

[3] S. Chen et al. Paragraph: a graph-based structural variant genotyper for short-read sequence data. *Genome Biology*, 20(1):291, 2019.

[4] H. P. Eggertsson et al. GraphTyper2 enables population-scale genotyping of structural variation using pangenome graphs. *Nature Communications*, 10(1):5402, 2019.

[5] G. Hickey et al. Genotyping structural variants in pangenome graphs using the vg toolkit. *Genome Biology*, 21(1):35, 2020.

[6] M. Rautiainen and T. Marschall. GraphAligner: rapid and versatile sequence-to-graph alignment. *Genome Biology*, 21(1):253, 2020.

Breakpoint detection in shallow and targeted panel in Rhabdomyosarcomas

Camille BENOIST[1], Stelly BALLET[2], Gaelle PIERRON[2], Victor RENAUT[1]

[1] Bioinformatique Clinique - PMDT, Institut Curie, 26 rue d'Ulm, 75005 Paris, Fran

[2] Unité de Génétique Somatique, Service d'oncogénétique, Institut Curie, Centre Hospitalier, 75005 Paris, France

Corresponding Author: camille.benoist@curie.fr

1. Introduction

Rhabdomyosarcoma (RMS) is the most common malignant mesenchymal tumor in children and adolescents. Based on clinico-pathologic features and genetic alterations, rhabdomyosarcomas are classified into embryonal, alveolar, spindle cell/sclerosing and pleomorphic subtypes. Each subtype shows distinctive morphology and has characteristic genetic abnormalities (variations, copy number alterations, gene fusions)[1].

In the CHEWIE project (Characterization of molecular Event betWeen dIagnosis and rElapse), we focused on the molecular alterations between diagnostic detection (RNAseq and WES) and relapse in pediatric embryonal and alveolar RMS (eRMS ans aRMS respectively). We followed biomarker (CNV, SNVs and fusions) in liquid biopsie (circulating tumor DNA) using two NGS approaches : low coverage WGS (shallow sequencing) and a dedicated panel targeting the PAX3-FOXO1 fusion characteristic of aRMS and the alterations of the RAS family genes found in eRMS.

Here we present the first results of the method we developed to identify the fusion breakpoints in shallow and targeted panel for 17 RMS samples. Based on bwa alignment, we extracted and analysed spanning reads (the two reads of the paired map in different location in the genome) and split reads (two portions of the same read map in different location in the genome). The combination of spanning and split reads metrics give us the precise location of breakpoint and their quantification.

Our tool is developed in python, and a git repository will be soon available allowing the identification and annotation of fusion breakpoints from NGS DNA library.

References

1. Agaram NP. Evolving classification of rhabdomyosarcoma. Histopathology. 2022 Jan;80(1):98-108. doi: 10.1111/his.14449. PMID: 34958505.

Title: A novel weight-based approach to determine cell type specificity from single cell datasets

Yanis Habtoun[1], *Kevin Cheeseman*[1], *Jean-Philippe Buffet*[1] & *Julien Cottineau*[1]

[1] WhiteLab Genomics, Future4Care, Watt-Biopark, 8 Rue Jean Antoine de Baïf, 75013 Paris

Corresponding Author: yhabtoun@whitelabgx.com

Abstract:

Haphazard integration of transgene sequences is the main blocking point in the generalization of gene therapy methods in medical settings. Although strict annotation of RNAseq data and identification of gene markers for tailored therapies could better restrain transgene integration towards targeted cells, most existing public databases featuring gene markers associated to SingleCell RNAseq and Bulk RNAseq data are not intended for gene therapy solutions. As such, we have built an atlas of biomarkers from 34 studies unifying bulk and single cell RNA-seq (n = 26 & n = 78 respectively) unifying alignments, normalization methods, annotations and metadata across databases storing bulk and single cell RNA-seq (SRA, HPA, GTEX, recount). Tissue annotations across the different databases were manually curated to create a general set of tissues and cell-types to be assigned to all datasets. All counts were obtained from alignment of fastq data with unified approached for bulk RNAseq and SingleCell RNAseq. Bulk datasets were normalized following current standards (TPM, pTPM). To account for cell type abundance for Single Cell data, an elaborate method for weight-scaling of expression data was developed. The weight-based method accounts for the differences in cell proportions between the cell types of different Single Cell Samples. The method is applied directly to normalized expression data whereas state of the art methods account for the proportion at the differential expression step. The newly computed proportions are further corrected with an enchancement of outliers and disenhancement of regular expression to make-up for the change in values without markedly affecting previous changes. Using the unified annotation of tissues, Bulk and Single Cell datasets were crosslinked to raise confidence-levels linked to gene markers. Cell types within Single Cell data were assigned using the AI approach of the CellAssign package, based on markers from public databases. Four specificity scores (SPM, JSS, TSI , Z-Score)[1] were tested and validated through data obtained from the HPA[2] database. Gene markers were obtained using the Seurat package, and then linked to specificity scores to construct a confidence-level beyond differential expression data.

Combining multiple steps from raw data towards gene markers, this weight-based approach leads to an accurate characterization of cell-type biomarkers applied for Gene Therapy from single cell RNA-seq datasets. Going forward, this approach will be used to integrate other OMICS datasets, e.g. ATAC-seq, ChiP-seq, and most notably OMICS data related to gene therapy experiments.

Keywords:

Gene Markers, Single Cell, Specificity Scoring, Gene therapy, Single Cell RNAseq

[1] Kryuchkova-Mostacci N. & Robinson-Rechavi M., 2016. A benchmark of gene expression tissue-specificity metrics. doi.org/10.1093/bib/bbw008

[2] Uhlén M., et al., 2015. Tissue-based map of the human proteome. DOI: 10.1126/science.1260419

Memory efficient subsampling strategy for large scale analysis of sequencing data

Timothé Rouzé, Antoine Lefevre, Caleb Smith, and Antoine Limasset
Univ. Lille, CNRS, Centrale Lille, UMR 9189 - CRIStAL, 59000 Lille, France

Corresponding author: antoine.limasset@univ-lille.fr

Since the beginning of high throughput sequencing technologies, a long-living challenge for bioinformaticians has been keeping up with the amount of generated data. With sequencers able to generate Terabytes of data each day for a constantly diminishing cost, the amount of data nucleotide on public databases is exploding. Comparing thousands of datasets containing up to billions of reads remains a scalability challenge and the focus of many methodological papers. The first step toward this direction was to avoid time-consuming alignment steps. Kmer comparisons methods showed good time performances while providing good similarity estimates [1]. However, to cost to represent billions kmers in memory quickly surpass most computers' available memory as a billion 32mer can represent a memory cost of 8GB and a billion 64mer 16GB. A solution is to rely on external memory [2] that is available in large amounts (commonly multiple TeraBytes). However, it dramatically slows down such analysis as Hard drives can be hundreds or thousands of times slower than the RAM. Moreover, the high usage of disks harms their durability. Another approach is to insert kmer in bloom filters [3] in order to divide the amount of memory used per kmer by order of magnitude at the price of some false positive rate. Studies showed that low false positive rates did not negatively impact downstream analysis. To achieve greater memory cost reduction, the only known technique is to apply some sub-sampling and select a fraction of kmer to index. Several methods were proposed to perform uniform sub-sampling with theoretical guarantee, modimizer/modminhah [4] , scaled minhash/FracMinHash [5] and showed their scalability on vast collections. In this work, we improve such schemes by combining them with the concept of superkmers [6]. Superkmers are a succession of overlapping kmer sharing a common subsequence called a minimizer. Such sequences can concisely represent dozen of kmer using less nucleotide than their plain representation. Due to these properties, superkmers have been used in several applications to reduce their memory usage. By applying sub-sampling directly on superkmers instead of kmers, we can benefit from the memory usage reduction granted by the superkmer usage. This way, we can either use less memory to represent the same amount of sub-sampled kmers or use comparable memory while indexing order of magnitude more kmers, thus improving the estimation accuracy at no cost. Moreover, superkmer usage can improve cache coherence and result in faster analysis. We apply our approach to several genomes and meta-genomes and show an order of magnitude improvement over state-of-the-art.

Acknowledgements

This work is financed by the ANR AGATE.

References

[1] Susana Vinga and Jonas Almeida. Alignment-free sequence comparison—a review. *Bioinformatics*, 19(4):513–523, 2003.

[2] Gaëtan Benoit, Pierre Peterlongo, Mahendra Mariadassou, Erwan Drezen, Sophie Schbath, Dominique Lavenier, and Claire Lemaitre. Multiple comparative metagenomics using multiset k-mer counting. *PeerJ Computer Science*, 2:e94, 2016.

[3] Nicolas Maillet, Claire Lemaitre, Rayan Chikhi, Dominique Lavenier, and Pierre Peterlongo. Compareads: comparing huge metagenomic experiments. In *BMC bioinformatics*, volume 13, pages 1–10. Springer, 2012.

[4] N Tessa Pierce, Luiz Irber, Taylor Reiter, Phillip Brooks, and C Titus Brown. Large-scale sequence comparisons with sourmash. *F1000Research*, 8, 2019.

[5] Luiz Carlos Irber, Phillip T Brooks, Taylor E Reiter, N Tessa Pierce-Ward, Mahmudur Rahman Hera, David Koslicki, and C Titus Brown. Lightweight compositional analysis of metagenomes with fracminhash and minimum metagenome covers. *bioRxiv*, 2022.

[6] Marek Kokot, Maciej Długosz, and Sebastian Deorowicz. Kmc 3: counting and manipulating k-mer statistics. *Bioinformatics*, 33(17):2759–2761, 2017.

AptaMat: a matrix-based algorithm to compare single-stranded oligonucleotides secondary structures

Thomas BINET[1], Bérangère AVALLE[1], Miraine DÁVILA FELIPE[2] and Irene MAFFUCCI[1]

[1] Université de technologie de Compiègne, UPJV, CNRS, Enzyme and Cell Engineering, Centre de recherche Royallieu - CS 60 319 - 60 203 Compiègne France

[2] Université de technologie de Compiègne, LMAC (Laboratory of Applied Mathematics of Compiègne), CS 60 319 - 60 203 Compiègne France

Corresponding Author: irene.maffucci@utc.fr

Single-stranded nucleic acids (ssNAs) are interesting molecules from both a biological and a biotechnological point of view. Indeed, they play important structural, functional and regulatory roles within the cell and, thanks to their ability in adopting specific conformations, they can bind to a large variety of molecular targets with high specificity and dissociation constants in the nano- to picomolar range. This makes both RNA and DNA oligonucleotides exploitable as therapeutic or diagnostic tools or as biosensors [1]. Therefore, since ssNAs function depends on their secondary and tertiary structures, the comparison of these two levels of arrangement can help to understand ssNAs roles and their interactions with other molecules, and to design ssNAs binding to a target of interest.

Here we focused on the first ssNAs level of organization, namely the secondary structure. So far, many algorithms aimed to compare ssNAs secondary structures have been developed [2]. However, to our knowledge, none of them allows at the same time an easy implementation, a straightforward results interpretation, and the ability of distinguishing between highly close structures. Therefore, we developed a matrix-based algorithm, called AptaMat, for the comparison and quantification of differences between structures of single stranded oligonucleotides of the same length (L). The algorithm takes as input two ssNAs secondary structures in the dot-bracket notation and creates for each of them a square matrix of $L \times L$, where each $(i,j)^{th}$ matrix entry is equal to either 1 or 0 if the nucleotides in positions i and j are paired or unpaired, respectively. The Manhattan distance is then used to find for each base pair of each structure the closest base pair in the other structure in a symmetric fashion. The distances between the closest pairs are summed up and normalized by the total number of base pairs in both structures.

We tested AptaMat on 10 ssNAs with known secondary structures: 5 taken from the work by Ivry et al. [3] work and 5 taken from the PDB database [4]. We compared AptaMat to the Hamming distance [5], RNAdistance [6] and to an image processing-based approach also based on matrices [3]. These implement algorithms belonging to 3 different class, namely character-based, tree-based and image processing-based, respectively, and they are commonly used for secondary structures comparisons. We showed that AptaMat can properly discriminate between different structures with a higher sensitivity as compared to the Hamming distance and RNAdistance. In addition, our method allows to more adequately rank the ssNAs structures as a function of their distance from a reference, which is not the case with the above-mentioned algorithms. Moreover, the results are easy to interpret with a reasonable threshold of 2 between close and far structures. Additionally, AptaMat is simple to implement and to manipulate. The python script implementing AptaMat is available on Github at https://github.com/GECgit/AptaMat.git.

References

1. P. K. Kulabhusan, B. Hussain, and M. Yüce, “Current perspectives on aptamers as diagnostic tools and therapeutic agents,” *Pharmaceutics*, vol. 12, no. 7. Multidisciplinary Digital Publishing Institute, pp. 1–23, Jul. 09, 2020
2. A. R. Gruber, S. H. Bernhart, I. L. Hofacker, and S. Washietl, “Strategies for measuring evolutionary conservation of RNA secondary structures.,” *BMC Bioinformatics*, vol. 9, p. 122, 2008
3. T. Ivry, S. Michal, A. Avihoo, G. Sapiro, and D. Barash, “An image processing approach to computing distances between RNA secondary structures dot plots,” *Algorithms Mol. Biol.*, vol. 4, pp. 1–19, 2009
4. H. M. Berman *et al.*, “The Protein Data Bank.,” *Acta Crystallogr. D. Biol. Crystallogr.*, vol. 58, no. Pt 6 No 1, pp. 899–907, Jun. 2002
5. R. W. Hamming, “Error Detecting and Error Correcting Codes,” *Bell Syst. Tech. J.*, vol. 29, no. 2, pp. 147–160, Apr. 1950
6. I. L. Hofacker, “Vienna RNA secondary structure server,” *Nucleic Acids Res.*, vol. 31, no. 13, pp. 3429–3431, Jul. 2003

Interpreting mass spectra differing from their peptide models by several modifications

Albane Lysiak[1,3], Guillaume Fertin[1], Géraldine Jean[1] and Dominique Tessier[2,3]
[1] Nantes Université, Ecole Centrale Nantes, CNRS, LS2N, UMR 6004, F-44000 Nantes, France
[2] INRAE, BIBS facility, F-44316, Nantes, France
[3] INRAE, UR BIA, F-44316, Nantes, France

Corresponding author: albane.lysiak@ls2n.fr

Introduction In proteomics, one of the main reasons for the poor identification rate of mass spectra when they are compared to candidate peptides (CP) is that most of them corresponds to the fragmentation of peptides carrying modifications. Open Modification Search (OMS) methods accept a wide range of mass difference within a Peptide-Spectrum Match (PSM) in order to improve the identification of spectra carrying modifications. Methods that were developped to identify and localize modifications in PSMs are either able to identify and localize a *single* modification, for instance SpecOMS [1] or MSFragger [2], or can combine several *a priori* known modifications to interpret a PSM, like PTMiner [3], or have a calculation time that is not compatible with the volume of data to process (e.g.[4]). Then, no efficient algorithm exists to interpret a PSM containing *several modifications* without *a priori*.

SpecGlob We developed `SpecGlob`, an algorithm that interprets PSMs by realigning a CP to its spectrum, even when *several* unknown modifications have occured. For each PSM, `SpecGlob` uses dynamic programming to determine the best alignment between a spectrum and its CP, while allowing the insertion of possibly multiple mass offsets. Given a PSM, `SpecGlob` outputs a sequence of amino acids interleaved by one or several mass offset(s), thus providing information on the modifications to apply to the CP so as to retrieve the spectrum sequence. Depending on the mass offsets values, the degree of difficulty to infer these modifications differs, something we quantified in order to evaluate the quality of `SpecGlob`. For example, if DYSIR plays the role of the experimental spectrum and DWYIR is the CP, the ouptut of `SpecGlob` allows us to infer two modifications, namely deletion of W and insertion of S. Hence we consider this PSM interpretation to be complete, because suggested mass offsets correspond to known (combinations of) amino acids masses.

Evaluation of SpecGlob Theoretical peptides from the human proteome (Ensembl 99) were compared to each other (self-identification excluded) using the SpecOMS software [1]. Resulting PSMs were then processed by `SpecGlob`, which takes as input masses of peaks of both spectra. Altogether, `SpecGlob` completely interprets many PSMs, even if they carry several scattered modifications. On the human dataset, `SpecGlob` returns a complete interpretation for roughly 30% of the 455,404 PSMs provided by SpecOMS. Our results also suggest that, even when a spectrum cannot be completely retrieved, a substantial portion of the initial amino acids sequence can still be determined.

Acknowledgements

Supported by the French National Research Agency (ANR-18-CE45-004), ANR DeepProt.

References

[1] M. David, G. Fertin, H. Rogniaux, and D. Tessier. SpecOMS: A Full Open Modification Search Method Performing All-to-All Spectra Comparisons within Minutes. *Journal of Proteome Research*, 16(8):3030–3038, August 2017. Publisher: American Chemical Society.

[2] A. Kong, F. Leprevost, D. Avtonomov, D. Mellacheruvu, and A. Nesvizhskii. MSFragger: ultrafast and comprehensive peptide identification in mass spectrometry-based proteomics. *Nature Methods*, 14(5):513–520, May 2017.

[3] Z. An, L. Zhai, W. Ying, X. Qian, F. Gong, M. Tan, and Y. Fu. PTMiner: Localization and Quality Control of Protein Modifications Detected in an Open Search and Its Application to Comprehensive Post-translational Modification Characterization in Human Proteome. *Molecular & Cellular Proteomics : MCP*, 18(2):391–405, February 2019.

[4] D. Tsur, S. Tanner, E. Zandi, V. Bafna, and P. Pevzner. Identification of post-translational modifications via blind search of mass-spectra. *Proceedings. IEEE Computational Systems Bioinformatics Conference*, pages 157–166, 2005.

MAM : Methylation Analysis of Microalgae.

Simon Brocard[1], Alexandre Cormier[2], Cyril Noël[2], Jérémy Berthelier[3], Grégory Carrier[1] and Rossana Sussarellu[1]

[1] IFREMER-PHYTOX-GenAlg, Centre Atlantique – rue de l'Ile d'Yeu,44980, Nantes, FRANCE

[2] IFREMER-SeBimer, Centre Bretagne – route de Sainte-Anne, 29280 Plouzané, FRANCE

[3] Institute of Science and Technology Graduate University (OIST), Okinawa, JAPON

Corresponding Author: Simon.Brocard@ifremer.fr

Microalgae are defined as unicellular or pluricellular undifferentiated organisms, eukaryotes or prokaryotes living in water. In some situations, phytoplankton proliferates, causing harmful algal blooms with considerable economic losses for aquaculture, fishing and tourism. [1]. Epigenetics plays an important role in the adaptation and proliferation processes of microalgae. Thanks to MinION Oxford Nanopore technology (ONT), which allows the sequencing of native DNA molecules without the need for bisulfite treatment, many tools for methylation detection have been developed by the scientific community. However, these algorithms are complicated to use for researchers with a background in biology and therefore require the implementation of automated, standardized and user-friendly solutions. In addition, the tools used for ONT are often updated and require regular monitoring to use them properly. Also, there is rarely information on the resources needed for the tool, tutorials explaining how the tool works, examples of analyses and test data sets. Finally, these tools are not designed for non-model organisms, but rather for mammalians or bacteria. Finally, most of the methylation analysis tools are based on a reference model but this model is rarely adapted to the species of interest.

In this context, the objective of the work is to build a reusable pipeline for the analysis of methylations in microalgae. For our project, we were able to design a model from Methylseq Illumina sequencing data of *Prymnesium parvum*. Between the multitude of tools are capable of identifying the positions of epigenetic modifications from Nanopore sequencing reads we chose two: Megalodon [2] and Deepsignal [3], [4]. Both tools work with a neural network and require a reference model to identify epigenetic modifications of sequences. We were able to compare the different results of these tools in order to select the "best" one according to the needs. In the future, it will be necessary to automate the data treatment processes in order to respect the FAIR principles (findable, accessible, interoperable and reusable).

References

[1] C. Belin et D. Soudant, « Trente années d'observation des microalgues et des toxines d'algues sur le littoral », janv. 2018, Consulté le: 20 janvier 2022. [En ligne]. Disponible sur: https://archimer.ifremer.fr/doc/00478/58981/

[2] *Megalodon*. Oxford Nanopore Technologies, 2022. Consulté le: 6 avril 2022. [En ligne]. Disponible sur: https://github.com/nanoporetech/megalodon

[3] P. Ni, *DeepSignal-plant*. 2022. Consulté le: 6 avril 2022. [En ligne]. Disponible sur: https://github.com/PengNi/deepsignal-plant

[4] P. Ni, *DeepSignal2*. 2022. Consulté le: 6 avril 2022. [En ligne]. Disponible sur: https://github.com/PengNi/deepsignal2

ToulligQC 2: fast and comprehensive quality control for Oxford Nanopore sequencing data

Karine Dias[1], Sophie Lemoine[1], Morgane Thomas-Chollier[1], Stéphane Le Crom[2,1], Médine Benchouaia[1], and Laurent Jourdren[1]

[1] GenomiqueENS, Institut de Biologie de l'ENS (IBENS), Département de biologie, École normale supérieure, CNRS, INSERM, Université PSL, 75005 Paris, France

[2] Sorbonne Université, CNRS, Institut de Biologie Paris-Seine (IBPS), Laboratory of Computational and Quantitative Biology (LCQB), F-75005, Paris

Corresponding Author: jourdren@bio.ens.psl.eu

The sequencing devices developed by Oxford Nanopore Technologies (ONT) produce long DNA sequence (> 200 kb) and full-length RNA. Sequencing and primary data acquisition are driven by the MinKNOW software, developed by ONT. MinKNOW stores the raw signal data as Fast5 files. Basecalling is then performed either during or after the acquisition step. Basecalling is usually achieved by the program Guppy, the official ONT basecaller. The output sequence reads are stored in FASTQ or Fast5 format. MinKNOW produces a Quality Control (QC) report as a PDF file at the end of the run. However this report only provides estimated information as it is based on non-basecalled and non-demultiplexed data. In addition, the metrics and scales that were provided by MinKNOW when we started RNA-seq applications in 2016 were not appropriate (unsuitable scales for RNA – which has been fixed since – and no barcode handling). It was thus necessary to develop a dedicated QC tool, flexible enough to handle both RNA and DNA sequencing.

The first version of ToulligQC is freely available since 2017, and used in production in our Genomics core Facility [1]. It allows users to quickly estimate the quality and homogeneity of their samples before running further analyses. Easy to use, this tool provides a detailed graphical output about the quality of Nanopore runs and exploratory data analysis, in the same spirit as the well-known FastQC program for short reads [2].

We introduce ToulligQC 2, a new major version of our QC software. ToulligQC 2 produces an improved HTML report with stylish and interactive plots obtained with the Plotly [3] library. The report contains exhaustive information about the sequencing run, basecalling and demultiplexing steps, such as: read count and length distributions, homogeneity of the run, location of potential flow cell spatial biases, statistics about pass and fail reads, PHRED score distribution and density distribution across read types, length/speed/quality and number of sequences over sequencing time, length/quality and read counts for each barcode. In addition to new graph types, all plots were qualitatively improved, and some of them provide alternative visualisation mode (e.g., boxplot and violin plot).

ToulligQC 2 has a reduced memory footprint and is faster (few minutes on a laptop) than the previous version. To facilitate interpretation of the graphs, each plot displays an “info” icon directly linking to the online help page on *GitHub* [4].

Because ONT protocols and bioinformatics tools are constantly evolving, ToulligQC 2 supports all versions of Guppy and the latest sequencing protocols. It can be used with all the Oxford Nanopore sequencing devices (MinION, GridION, PromethION), and remains compatible with both 1D and $1D^2$ chemistries. It takes as input the sequencing summary file generated by the Guppy basecaller and the sequencing telemetry file, if available.

ToulligQC 2 is an *open source* software published under GPL3 and CeCILL licences. It can be freely downloaded on *Github* [4], as a *Docker image* (genomiquepariscentre/toulligqc) [5], and as a *PyPy package* [6].

Acknowledgements

The IBENS genomics core facility 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] https://genomique.biologie.ens.fr/
[2] https://www.bioinformatics.babraham.ac.uk/projects/fastqc/
[3] https://plot.ly
[4] https://github.com/GenomiqueENS/toulligQC
[5] https://hub.docker.com/r/genomicpariscentre/toulligqc
[6] https://pypi.org/project/toulligqc/

Benchmark of Nvidia Clara Parabricks suite for GPU-accelerated bioinformatic processing of raw RNA-seq data

Etienne BARDET[1], Pauline BAZELLE[1], Christophe BATTAIL[1] and KATY CONSORTIUM[2]

[1] Laboratoire Biologie et Biotechnologies pour la Santé, IRIG, UMR 1292 INSERM-CEA-UGA, Univ. Grenoble Alpes, 38000 Grenoble, France.

[2] https://katy-project.eu, European Unions Horizon 2020 research and innovation programme, Grant agreement No 101017453

Corresponding Author: christophe.battail@cea.fr

The emergence of personalized medicine requires being able to produce and process huge amounts of biological data generated from patients' biological samples, quickly and at a reasonable cost. Modern sequencing techniques make it possible to obtain a large amount of genetic data about a patient relatively quickly, but bioinformatics processing of this data takes time. One of the ways to accelerate these bioinformatics analyses would be to use the computing power of GPUs, a strategy commonly used for applications using AI methods. However, this practice is not uniformely spread throughout all of the bioinformatics fields. There have been so far few tools dedicated to omics data analyses accelerated on GPUs, producing results not perfectly reproducible with those generated by the original implementations on CPUs [1].

NVIDIA, one of the world leaders in GPU production, recently released the 3rd version of its Clara Parabrick suite. This software suite accelerates, by using GPUs, popular bioinformatics tools for genomics and transcriptomics data analysis available in open source. Although the approach is interesting and welcome, their tools have been the subject of only relatively limited studies about the promised performance gain. A previous publication focused on Parabrick's ability to process genomics data [2], but no benchmark so far have been done for RNA-seq data. Our project aims to measure the gains in computation time, memory footprint and energy consumption between GPU-accelerated tools and original CPU tools used in transcriptomics data analysis. We will also make sure to assess the level of reproducibility between the results generated by the GPU-accelerated tools, compared to the original versions.

References

[1] Yongchao Liu and Bertil Schmidt. *CUSHAW: a CUDA compatible short read aligner to large genomes based on the Burrows–Wheeler transform.* Bioinformatics, Volume 28, Issue 14: 1830–1837, 2012

[2] Karl R. Franke and Erin L. Crowgey. *Accelerating next generation sequencing data analysis: an evaluation of optimized best practices for Genome Analysis Toolkit algorithms*. Genomics & Informatics, 18(1): e10, 2020

A nextflow workflow for peptide sequence design in a targeted proteomic approach

Sylvère BASTIEN[1,2], Pauline FRANÇOIS[1,2], Iulia MACAVEI[3], Karen MOREAU[1], Jérôme LEMOINE[3] and François VANDENESCH[1,2]

[1] CIRI, Centre International de Recherche en Infectiologie, Université de Lyon, Inserm, U1111, Université Claude Bernard Lyon 1, CNRS, UMR5308, ENS de Lyon, Lyon, France

[2] Hospices civils de Lyon, RHU IDBIORIV, Institut des Agents Infectieux, Lyon, France

[3] Univ Lyon, CNRS, Université Claude Bernard Lyon 1, Institut des Sciences Analytiques, UMR 5280, 5 rue de la Doua, F-69100 Villeurbanne, France

Corresponding Author: francois.vandenesch@univ-lyon1.fr

An important step in using targeted proteomic approaches is to define the peptide sequences that allow the detection of all variants of a protein of interest. Sequence variability can be extremely high, especially in bacteria, due to their rate of evolution and ability to adapt to their environment. Another important issue in the choice of peptide sequences is the possibility of targeting a protein whose coding region has experienced a duplication event in its history. It is therefore crucial when quantifying a protein that the designed peptides are both specific to the protein of interest while not being found in repeated regions whose copy number may vary from one genome to another.

In this context, we have developed a nextflow [1] pipeline to define a minimal list of peptide sequences that can be used in targeted mass spectrometry to detect any variant of a protein of interest. The first optional part consists in the constitution of a non-redundant protein database via CD-HIT [2] from all the protein sequences available on ncbi via a user-defined taxid. The second part corresponds to the selection of the set of variants to be retained according to their frequency (user-defined threshold) in order to be able to constitute a set of protein sequences necessary to establish the final list of peptide sequences. This pipeline is based on the sequential use of several alignment tools both local (BLASTP [3]) and global (EMBOSS [4]) as well as several filters focusing on best match, percentage of mismatches, insertions/deletions and coverage. The methodology for creating the peptide sequence list is a computer simulation of the use of in-vitro trypsin hydrolysis. The user has the possibility to modify this global peptide sequence list according to his own defined peptide characteristics. Finally, a minimum list of peptide sequences is established according to a threshold entered by the user corresponding to the minimum number of times that each variant must be covered. The ultimate step consists of validating the final peptides, notably by aligning these peptides with the initial database.

This pipeline was used to select peptide sequences for the quantification of 44 *Staphylococcus aureus* virulence proteins in a targeted proteomics approach using high-throughput mass spectrometry.

References

[1] P. Di Tommaso, M. Chatzou, E. W. Floden, P. P. Barja, E. Palumbo, and C. Notredame, 'Nextflow enables reproducible computational workflows', *Nat Biotechnol*, vol. 35, no. 4, pp. 316–319, Apr. 2017, doi: 10.1038/nbt.3820.

[2] L. Fu, B. Niu, Z. Zhu, S. Wu, and W. Li, 'CD-HIT: accelerated for clustering the next-generation sequencing data', *Bioinformatics*, vol. 28, no. 23, pp. 3150–3152, Dec. 2012, doi: 10.1093/bioinformatics/bts565.

[3] 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, Art. no. 3, Oct. 1990, doi: 10.1016/S0022-2836(05)80360-2.

[4] P. Rice, I. Longden, and A. Bleasby, 'EMBOSS: The European Molecular Biology Open Software Suite', *Trends in Genetics*, vol. 16, no. 6, pp. 276–277, Jun. 2000, doi: 10.1016/S0168-9525(00)02024-2.

MetExploreViz: Visualization tool for metabolic networks

Jean-Clément Gallardo[1], Maxime Chazalviel[2], Nathalie Poupin[1], Clément Frainay[1], Ludovic Cottret[3], Florence Vinson[1] and Fabien Jourdan[1]

[1] Toxalim, Université de Toulouse, INRAE, ENVT, INP-Purpan, UPS, Toulouse, France
[2] MedDay Pharmaceuticals, Paris, France
[3] LIMP, Université de Toulouse, INRAE, CNRS, Castanet-Tolosan, France

Corresponding Author: jean-clement.gallardo@inrae.fr

1 Context

The study of metabolism is a complex but key research focus to develop a better understanding of living systems. Over time, more and more omics data has been generated and collected, providing researchers with large datasets. Furthermore, because the study of life is an ever-evolving subject, scientists are constantly updating their knowledge and thus metabolic networks. As a result, we are left with more massive and complicated networks.

Exploration of metabolic data can be done on a small scale, with the study of an isolated metabolic pathway to understand the degradation or synthesis of a specific molecule. But this exploration can also be done on a larger scale, with the analysis of several interconnected metabolic pathways for instance to understand the metabolic disturbances induced by a specific condition on the whole organism, tissue or cell for instance. In each of these situations, graph visualization is advantageous for studying and understanding the data, by smooth navigation of relationships that links compounds concentrations through reaction chains.

We develop a tool named MetExploreViz [1], a visualization and exploration tool for metabolic networks, that intends to provide users with a large panel of features for automatic drawing, graph style management and multi-scale data exploration.

2 MetExploreViz

MetExploreViz is an open source web component that can be easily integrated into other projects or websites. It provides features dedicated to the visualization of metabolic networks and pathways, making it a versatile tool for analyzing omics data in a biochemical context.

The MetExploreViz view will allow the users to manage the drawing and produce publication-ready figures. For example, it is possible to manage the rendering of the network by modifying the representation of the nodes, links and their positions. This can be done manually or by using implemented drawing algorithms such as the popular force algorithm, a circular drawing algorithm or a hierarchical layout.

MetExploreViz-based apps also offer to conduct further analysis on users' datasets: it is possible to map metabolites or reactions on the whole network or specific pathways, import and map a list of side compounds that will be removed to lighten the drawing, extract sub-networks, and more.

Recently, new features have been developed to enhance data interpretation. It is now possible to visualize flux modeling results through the network links. This feature includes importing flux data, mapping it to the corresponding reactions, and finally visualizing it. It is also possible to perform comparative analysis of two conditions and to display standard deviation values.

Another feature that will be available soon, will allow users to reconstruct a sub-network step by step from their data, with assistance of a recommender system [2]. This feature, the Network Explorer, will allow to move from a large network to a sub-network of interest in order to extract relevant information.

References

[1] Chazalviel M., Frainay C., Poupin N., Vinson F., Merlet B., Gloaguen Y., Cottret L., and Jourdan F. MetExploreViz: web component for interactive metabolic network visualization. *Bioinformatics*, 34(2):312–313, jan 2018.

[2] Frainay C., Aros S., Chazalviel M., Garcia T., Vinson F., Weiss N., Colsch B., Sedel F., Thabut D., Junot C., and Jourdan F. MetaboRank: network-based recommendation system to interpret and enrich metabolics results. *Bioinformatics*, 35(2), 274-283, 2019.

Improving attribute exploration for the detection and correction of anomalies in an agroecological knowledge base

Nassif Saab[1], Marianne Huchard[1] and Pierre Martin[2]
[1] LIRMM, Université de Montpellier, 161 rue Ada, 34095, Montpellier, France
[2] CIRAD, UPR AIDA, Avenue Agropolis, 34398, Montpellier, France

Corresponding author: nassif.saab@lirmm.fr

Data cleaning is crucial to the knowledge discovery process. Knowledge bases such as Knomana [1] rely on data wrangling to standardise and subsequently centralise information extracted from multiple sources. This makes Knomana prone to anomalies, i.e. to incorrect or incomplete descriptions of plant use, which may cause its users to draw wrong conclusions during knowledge discovery. To detect and correct these anomalies, we propose using Attribute Exploration (AE) [2] to acquire expert knowledge and apply it to identify anomalies and correct or complete the descriptions. It is a process of Formal Concept Analysis, which considers data tables describing binary relationships between objects and attributes. AE relies on the computation of the Duquenne-Guigues basis, a complete, consistent and nonredundant set of implication rules, i.e. regularities of the form "if there is X, then there is always Y" [3]. The expert is asked to validate the generated implications or provide a counterexample when an invalid rule is presented. Tools like ConExp [4] implement AE. With Knomana holding 35 attributes covering over 45,000 descriptions of plant use, the number of computed rules is in the thousands [5]. Therefore, it is consequential to have a pertinent and time-saving order of displaying these rules.

To tackle the problem at hand, this poster presents an improvement of AE. During AE, the computed rules are consecutively shown to the expert in the lectic order, where set A is presented before set B if the smallest differing element belongs to B. According to this definition, the lectic order does not consider the nature of the data it is addressing, and consequently, the implications are not displayed in a meaningful order, i.e. an order that regards the expert's interest in a particular type of data. Thereupon, we propose that experts sort the data prior to exploring the attributes. By providing experts with the means to group attributes into categories and order them by relevance, table columns are rearranged in conformity with the definition of the lectic order for the purpose of generating the most relevant implications first. Applying this change to a single data table allowed to accommodate AE to the interests of the expert. As a next step, we plan to extend this technique to relational data to render it applicable to datasets that employ ternary relationships, as is the case in the agroecological knowledge base Knomana.

Acknowledgements

This work is supported by Montpellier University KIM (Key Initiatives MUSE) DATA & LIFE SCIENCES through an interdisciplinary internship grant.

References

[1] Pierre J. Silvie, Pierre Martin, Marianne Huchard, Priscilla Keip, Alain Gutierrez, and Samira Sarter. Prototyping a knowledge-based system to identify botanical extracts for plant health in sub-saharan africa. *Plants*, 10(5), 2021.

[2] Bernhard Ganter and Sergei Obiedkov. *Conceptual Exploration*. Springer Berlin Heidelberg, 2016.

[3] Alexandre Bazin and Jean-Gabriel Ganascia. Computing the Duquenne–Guigues basis: an algorithm for choosing the order. *International Journal of General Systems*, 45(2):57–85, September 2015.

[4] Serhiy A. Yevtushenko. Conexp, 2022.

[5] Lina Mahrach, Alain Gutierrez, Marianne Huchard, Priscilla Keip, Pascal Marnotte, Pierre Silvie, and Pierre Martin. Combining implications and conceptual analysis to learn from a pesticidal plant knowledge base. In *Graph-Based Representation and Reasoning*, pages 57–72. Springer International Publishing, 2021.

MyGOD : Interface de visualisation et d'analyse de données provenant des observatoires génomiques marins

Charlotte Andre[1], Mark Hoebeke[1], Nicolas Henry[1], Cyril Noel[2], Patrick Durand[2] and Erwan Corre[1]

[1] CNRS - Sorbonne Université - Plateforme ABIMS - Station Biologique de Roscoff, Place Georges Teissier, 29680, Roscoff, France
[2]IFREMER-IRSI-Service de Bioinformatique (SeBiMER), Centre Bretagne, 1625 Route de Sainte-Anne, 29280 Plouzané, France.

Corresponding Author: charlotte.andre@sb-roscoff.fr

Un observatoire génomique marin est un dispositif d'acquisition et de bancarisation de données hétérogènes et complexes (données génomiques, comptages d'espèces, données physico chimiques), centré sur une zone géographique réduite et dont le principal objectif est de réaliser un suivi à long terme du fonctionnement et de l'évolution d'un écosystème marin. Actuellement, plusieurs initiatives de mise en place d'observatoires génomiques marins sont en cours dans les régions Bretagne - Pays de Loire.

Le projet MyGOD *(Manipulate your Genomics Observatory Data)*, financé par la région Bretagne dans le cadre du réseau Biogenouest, vise à aider les équipes impliquées dans l'exploitation de données issues de ces observatoires génomiques, à explorer et interpréter ces données. Ceci, en leur fournissant un outil de visualisation intégré et ergonomique : MyGOD principalement développé par la plateforme ABiMS et dont les fonctionnalités s'inspirent de celles de OBA (Ocean Barcode Atlas), développé par le MIO dans le cadre du projet Tara Ocean , s'appuie également sur des composants de visualisation réalisés en R par le SEBIMER dans le cadre du projet de pipeline d'analyse de données SAMBA. A terme, MyGOD vise également à permettre à un public plus large d'accéder à des visualisations de phénomènes remarquables ("success stories") qui auront été composées et rendues publiques par les experts des jeux de données. Nous avons par ailleurs travaillé sur l'automatisation de l'intégration des données, la création d'un site Web combinant les technologies Django (Python) et Vue.js (Javascript) et la mise en place de graphiques dynamiques en D3.js permettant la mise en corrélation des données. La nature différente des jeux de données candidats à l'intégration dans MyGOD a par ailleurs permis d'entamer une réflexion sur la standardisation des leurs formats ainsi que des métadonnées qui les accompagnent.

Ce travail est le fruit d'une collaboration entre la plateforme ABiMS, les équipes locales de la station Biologique de Roscoff, le Sebimer IFREMER et l'UBO (Brest), le LS2N (Nantes), le MIO (Marseille) et le Genoscope (Evry).

The Phaeoexplorer Database: a Multi-Scale Genomic and Transcriptomic Data Resource for the Brown Algae

Loraine Brillet-Guéguen[1,2], Arthur Le Bars[3,2], Romain Dallet[2], Gildas Le Corguillé[2], Olivier Godfroy[1], J. Mark Cock[1], Susana M. Coelho[1,4] and Erwan Corre[2]

[1] Sorbonne Université, CNRS, Integrative Biology of Marine Models (LBI2M), Station Biologique de Roscoff (SBR), 29680, Roscoff, France

[2] CNRS, Sorbonne Université, FR2424, ABiMS, Station Biologique, 29680, Roscoff, France

[3] CNRS, Institut Français de Bioinformatique, IFB-core, UMS 3601, Évry, France

[4] Max Planck Institute for Biology, Department of Algal Development and Evolution, Tübingen, Germany

Corresponding Author: loraine.gueguen@sb-roscoff.fr

Abstract

The Phaeoexplorer project aims to generate annotated genome assemblies and transcriptome data for a broad range of brown algal species to address key questions about their biology and evolutionary history. More than 60 genomes of brown algae and closely-related sister species have been sequenced to date.

To provide the community with a collaborative hub for accessing, visualizing and analyzing the brown algal genome and transcriptome resources, we have developed a web portal using the Django framework (https://phaeoexplorer.sb-roscoff.fr) to house the annotated genome sequences along with a broad range of associated resources. These resources include an integrated environment based on the Galaxy Genome Annotation project dedicated to the management and visualization of genomic data through user-friendly interfaces (including JBrowse genome browsers), deployed in an automated way with a set of custom Python tools [1] (https://gitlab.sb-roscoff.fr/abims/e-infra/gga_load_data). Other resources include information about the sequenced strains; assembly and annotation metrics; data download facilities; SequenceServer [2] BLAST facilities, deployed with an Ansible role (https://galaxy.ansible.com/abims_sbr/sequenceserver) and a R Shiny web application designed to explore RNAseq data of the model alga *Ectocarpus* (https://rnaseqaggregator.sb-roscoff.fr/ectocarpus). Over the next few months, we plan to extend the Phaeoexplorer database with additional resources (experimental protocols, genomes of associated bacterial symbionts and a Genomicus-based [3] comparative genomics resource). Still partially restricted, in the long term, the objective is for the Phaeoexplorer database to be a user-friendly public access point to brown algal genomes for the entire phycology community, with regular genome releases.

In the future, similar databases will be implemented for red algae and fungi; and, within the context of the European Reference Genome Atlas project and other future large genome sequencing programs, in partnership with the BIPAA/Genouest and the SeBiMER/Ifremer bioinformatics platforms, we plan to further automate and scale up the omics data integration pipeline.

Acknowledgements

This work is supported by the ANR programs France Génomique Phaeoexplorer (ANR-10-INBS-09), PIA IDEALG (ANR–10–BTBR–04) and IFB (ANR-11-INBS-0013), by the ERC program SEXSEA (grant ID 638240) and by the Biogenouest network. We thank the Genoscope for providing the sequence assembly and the structural annotation of the Phaeoexplorer genomes; Agnieszka P. Lipinska (Max Planck Institute) for primary data analysis.

References

1. Brillet-Guéguen L, Le Bars A *et al.* Automated Deployment of Genome Databases for Marine Brown Algae Using Galaxy Genome Annotation. *F1000Research* 2021, 10:596, doi: 10.7490/f1000research.1118615.1 (poster)
2. Anurag Priyam *et al*. Sequenceserver: a modern graphical user interface for custom BLAST databases. *Molecular Biology and Evolution*, Vol. 36, Issue 12, 2019, Pages 2922–2924, doi: 10.1093/molbev/msz185
3. Nga Thi Thuy Nguyen *et al*, Genomicus in 2022: comparative tools for thousands of genomes and reconstructed ancestors, *Nucleic Acids Research*, Vol. 50, Issue D1, 2022, Pages D1025–D1031, doi: 10.1093/nar/gkab1091

Accessibilité des modèles pharmacocinétiques physiologiques

Sidonie Halluin[1] et Cléo Tebby[1]

[1] Institut national de l'environnement industriel et des risques (INERIS), Parc Technologique Alata, BP N°2, 60550 Verneuil-en-Halatte, France

Auteur correspondant: halluin.sidonie@gmail.com

Les modèles pharmacocinétiques physiologiques (PBPK) sont des modèles mathématiques qui prédisent l'absorption, la distribution, le métabolisme et l'excrétion de substances chimiques chez un organisme. Celui-ci est représenté par un ensemble de compartiments liés entre eux par le compartiment sanguin. Ce type de modèle est utilisé pour évaluer l'exposition interne d'un organisme d'après son exposition externe à une substance, notamment dans le cadre de réglementations sanitaires.

Ce travail porte sur les modèles PBPK développés par l'institut national de l'environnement industriel et des risques. L'unité toxicologie expérimentale et modélisation a publié en 2010 un modèle PBPK de référence[1], qui suit un organisme humain durant sa vie entière. Depuis, de nombreuses versions ont été développées pour s'adapter aux différents comportements complexes des substances. Ces modèles sont codés en GNU MCSim, qui est un langage basé sur le langage C. L'objectif de ce travail est de rendre l'utilisation des modèles PBPK plus accessible pour les chercheurs ainsi que pour les agences réglementaires.

Le premier axe est la création d'un modèle PBPK unique et applicable à un maximum de substances. Pour ce faire, les différentes versions ont été collectées, mises à jour et comparées entre elles. Le modèle générique réunissant les différentes versions est en cours de développement et sera mis à disposition sur GitLab.

Le second axe est le développement d'une interface pour représenter graphiquement la dynamique des prédictions du modèle. Elle permet de prédire l'évolution temporelle des concentrations internes à partir de scénarios d'exposition définis par l'utilisateur. Cette interface génère trois outils de visualisation des résultats, de façon automatique et paramétrable :

- Un tableau ainsi qu'un graphique des concentrations internes d'une substance dans chaque compartiment

- Un schéma du corps humain où les organes changent de couleur en fonction de leur concentration interne

L'interface graphique est en cours de développement et est codée en R Shiny. L'animation du corps humain est réalisée via un fichier R qui génère du code CSS de SVG. Cet outil, destiné à faciliter l'obtention et l'interprétation des résultats des modèles PBPK, aidera à l'évaluation des expositions aux substances chimiques en matière de santé humaine.

Mots clés

Relation dose-effet

Toxicocinétique

Modèle pharmacocinétique physiologique (PBPK)

Animation SVG

Interface graphique R Shiny

Remerciements

Je souhaite remercier l'unité toxicologie expérimentale et modélisation (TEAM) de l'institut national de l'environnement industriel et des risques (INERIS) pour m'avoir accueilli dans le cadre de ce travail en alternance.

J'aimerais remercier particulièrement Mme Cléo Tebby, ingénieure d'étude et de recherche de l'unité TEAM. En tant qu'encadrante, elle m'a accordé beaucoup de temps pour me former, répondre à mes questions et me conseiller.

Je remercie également Mme Céline Brochot, responsable de l'unité TEAM, pour m'avoir permise de participer aux journées ouvertes en biologie, informatique et mathématiques 2022.

Enfin, je tiens à remercier l'équipe pédagogique du master de bioinformatique de l'université Rouen Normandie, pour leur enseignement et leur accompagnement tout au long de cette alternance.

Références

1. Rémy Beaudouin, Sandrine Micallef et Céline Brochot, 2010, *A stochastic whole-body physiologically based pharmacokinetic model to assess the impact of inter-individual variability on tissue dosimetry over the human lifespan*, Regulatory Toxicology and Pharmacology. https://doi.org/10.1016/j.yrtph.2010.01.005

Omnicrobe, an open-access database of microbial habitats, phenotypes and uses extracted from text

Sandra DEROZIER[1], Robert BOSSY[1], Louise DELEGER[1], Mouhamadou BA[1,2], Estelle CHAIX[1], Valentin LOUX[1,2], Hélène FALENTIN[3], Claire NEDELLEC[1]

[1] Université Paris-Saclay, INRAE, MaIAGE, Jouy-en-Josas, France

[2] Université Paris-Saclay, INRAE, BioinfOmics, MIGALE Bioinformatics Facility, Jouy-en-Josas 78350, France

[3] INRAE, STLO, Rennes, France

Corresponding Author: sandra.derozier@inrae.fr

The drastic increase in microbe descriptions, habitats, phenotypes and uses in databases, reports and papers presents a two-fold challenge for the access to the information. The integration of heterogeneous data requires a standardized representation and the normalization of textual descriptions by semantic analysis. Recent information extraction technologies from the text mining domain offer a powerful way to detect and structure textual information along ontology-based representations.

The Omnicrobe application (https://omnicrobe.migale.inrae.fr) uses an Information Extraction workflow to populate its database. The workflow is designed to (1) extract microorganism taxa, their habitats, their phenotypes and their uses and (2) categorize the extracted information with taxa from the NCBI (National Center for Biotechnology Information) taxonomy [1] and concepts from the OntoBiotope ontology [2]. The Omnicrobe database contains around 1 million descriptions of microbe properties that are created by analyzing and combining six information sources, i.e. biological resource catalogues (e. g. INRAE CIRM, DSMZ through BacDive [3]), sequence database (GenBank) and scientific literature (PubMed abstracts).

Omnicrobe offers powerful ways to express simple and complex ontology-based queries to support studies in various domains of microbiology. Omnicrobe also exposes an API (Application Programming Interface) that allows users to automatically integrate microbe biodiversity knowledge in external information systems. The use of Omnicrobe to quickly target useful strains in a food innovation application [4] illustrates how it can provide an easy-to-use support in the resolution of scientific questions related to the habitats, phenotypes and uses of microbes.

Acknowledgements

The authors thanks the Migale platform for providing the resources to run Omnicrobe services (MIGALE, INRAE, 2020. Migale bioinformatics Facility, doi: 10.15454/1.5572390655343293E12).

References

1. Schoch CL, Ciufo S, Domrachev M, Hotton CL, Kannan S, Khovanskaya R, Leipe D, Mcveigh R, O'Neill K, Robbertse B, Sharma S, Soussov V, Sullivan JP, Sun L, Turner S, Karsch-Mizrachi I. NCBI Taxonomy: a comprehensive update on curation, resources and tools. *Database (Oxford)*, 2020.
2. Chaix E., Deléger L., Bossy R., Nédellec C. Text mining tools for extracting information about microbial biodiversity in food. *Food Microbiology*, 2018.
3. Reimer LC, Vetcininova A, Carbasse JS, Söhngen C, Gleim D, Ebeling C, Overmann J. BacDive in 2019: bacterial phenotypic data for High-throughput biodiversity analysis. *Nucleic Acids Res.,* 47(D1):D631-D636, 2019.
4. Harlé O., Falentin H., Niay J., Valence F., Courselaud C., Chuat V., Maillard M-B., Guédon E., Deutsch S-M., Thierry A. Diversity of the metabolic profiles of a broad range of lactic acid bacteria in soy juice fermentation. *Food microbiology*, *89*, 2020.

IDy-Path : Identification Dynamique des épidémies de Pathogènes

Bryce Leterrier[1] and Meriadeg Le Gouil[1]

[1] CHU de Caen/Laboratoire de Virologie, Avenue Georges Clemenceau, 14000, Caen, France

Corresponding Author: bryce.leterrier@gmail.com

Résumé

Les épidémies virales représentent un enjeu de santé publique majeur. Un système de détection et de contrôle est donc mis en place notamment au sein des structures hospitalières. Certains virus respiratoires, en particulier ceux présentant des caractéristiques telles qu'un taux de transmission élevé ou un mode de vie saisonnier [1] en font des cibles privilégiées de surveillance. De même, dans le contexte de virus émergents comme le SARS-CoV 2 (Servere Acute Respiratory Syndrome Coronavirus 2) et pour le suivi de la pandémie qu'il a engendré, la mise à disposition des données au plus près de la réalité apparaît donc nécessaire pour apporter une réponse médicale plus efficace mais aussi pour permettre une meilleure gestion des épidémies.

Le projet IDy-Path a vocation à fournir une interface web R-Shiny [2] exploitant la base de données de l'hôpital et accessible via le réseau interne au personnel hospitalier. Elle permet la visualisation en temps réel du suivi d'une dizaine d'espèces virales respiratoires. Les données disponibles recensent les prévalences depuis 2015 afin de permettre une analyse épidémiologique rétrospective et qui peuvent servir de base d'apprentissage pour un module de prédiction des départs d'épidémies.

Un module spécifique a été développé pour le virus de la rougeole dans le cadre de l'activité du CNR ROR (Centre National de Référence Rougeole-Oreillons-Rubéole) qui est porté par le CHU de Caen. Il permet des analyses plus précises concernant l'origine physiologique et la nature des échantillons traités, les analyses génétiques effectuées sur ces derniers ainsi que diverses représentations géographiques de leurs provenances.

L'ensemble des représentations générées fournit un outil supplémentaire d'aide au diagnostic guidé par la situation épidémique en temps réel. L'anticipation de l'entrée en phase épidémique des virus saisonniers permet aussi d'aider à la gestion des services en prévision d'une forte circulation virale synonyme d'augmentation de l'activité hospitalière.

References

[1] M. Moriyama, W. J. Hugentobler, et A. Iwasaki, « Seasonality of Respiratory Viral Infections », *Annu Rev Virol*, vol. 7, nº 1, p. 83-101, sept. 2020, doi: 10.1146/annurev-virology-012420-022445.

[2] W. Chang *et al.*, « shiny: Web Application Framework for R. R package version 1.7.1. », 2021. https://CRAN.R-project.org/package=shiny

The advantage of having a web dedicated group in a Bioinformatic team

Rachel TORCHET[1], Bryan BRANCOTTE[1], Hippolyte KENGNI[2], Lucie LAMOTHE[2], Fabien MAREUIL[1], Remi PLANEL[1], Herve MENAGER[1]

[1] Institut Pasteur, Université Paris Cité, Bioinformatics and Biostatistics Hub, F-75015 Paris, France

[2] Institut Français de Bioinformatique, France

Corresponding Author: rachel.torchet@pasteur.fr

The WINTER group is a software development team focusing mainly on Web technologies for publishing and sharing scientific tools, analysis, data and workflows. We provide our expertise to the scientists of the Institut Pasteur campus, covering a broad range of services and expertise to design, develop, maintain, and publish software tools and databases on the Web. As part of the Hub mission, our projects cover a wide variety of scientific topics (Structural Bioinformatics, Transcriptomics, Statistical Genetics, etc.).

Over the past few years, our group has created more than 15 web applications and databases, in collaboration with research units and other groups of the Hub, and with the support of the IT department, including for instance:

- iPPI-DB: a database of modulators of protein-protein interactions [1]. The data are retrieved from the literature either peer reviewed scientific articles or world patents. A large variety of data is stored within iPPI-DB: structural, pharmacological, binding and activity profile, pharmacokinetic and cytotoxicity when available, as well as some data about the PPI targets themselves.
- CRISPR-browser: a genome browser to visualize the results of CRISPR-dCas9 screens in bacteria [2]. You can upload your own data or navigate through published datasets.
- Modelisation-COVID: an information website that publishes the work carried out by the Mathematical Modelling of Infectious Diseases Unit on COVID-19 [3].
- ARIAweb: a software for automated NOE assignment and NMR structure calculation [4]. This specialized application accesses the computing infrastructure through the Galaxy API.

The WINTER group also acts as a transversal support to the different thematic groups of the Hub, providing a high level of expertise to support bioinformaticians and biostatistician with their web application development. In this way, having a web dedicated group in a bioinformatic team appears as a great opportunity to bring software development at the foreground of the bioinformatic research.

References

1. Rachel Torchet, Karen Druart, Luis Checa Ruano, Alexandra Moine-Franel, Hélène Borges, Olivia Doppelt-Azeroual, Bryan Brancotte, Fabien Mareuil, Michael Nilges, Hervé Ménager, Olivier Sperandio, *The iPPI-DB initiative: a community-centered database of protein–protein interaction modulators*, Bioinformatics, Volume 37, Issue 1, 1 January 2021
2. Alicia Calvo-Villamañán, Jérome Wong Ng, Rémi Planel, Hervé Ménager, Arthur Chen, Lun Cui, David Bikard, *On-target activity predictions enable improved CRISPR–dCas9 screens in bacteria*, Nucleic Acids Research, Volume 48, Issue 11, 19 June 2020.
3. Hozé N, Paireau J, Lapidus N, Tran Kiem C, Salje H, Severi G, Touvier M, Zins M, de Lamballerie X, Lévy-Bruhl D, Carrat F, Cauchemez S. *Monitoring the proportion of the population infected by SARS-CoV-2 using age-stratified hospitalisation and serological data: a modelling study.* Lancet Public Health. June 2021.
4. Allain F, Mareuil F, Ménager H, Nilges M, Bardiaux B. *ARIAweb: a server for automated NMR structure calculation.* Nucleic Acids Research. 2020 Jul 2;48(W1):W41-W47

Circscan: a shiny application to decipher circRNA-miRNA-mRNA networks from multi-level transcriptomic data.

Rose-Marie Fraboulet[1], Yanis Si Ahmed[1], Sébastien Corre[1], Marie-Dominique Galibert[1,2], Yuna Blum[1]

[1] Univ Rennes, CNRS, INSERM, IGDR (Institut de Génétique et Développement de Rennes) - UMR 6290, ERL U1305, F-35000 Rennes, France.

2 Department of Molecular Genetics and Genomics, Hospital University of Rennes (CHU Rennes), F-35000 Rennes, France.

Corresponding Author: yuna.blum@univ-rennes1.fr

Abstract

Non-coding RNAs have been poorly studied until now, while representing a large part of the human transcriptome and having an important role in cancer [1]. Among these, circular RNAs (circRNAs) have recently been discovered for their microRNA (miRNA) sponge function [2], which allows them to modulate the expression of miRNA target genes: they then take on the role of competitive endogenous RNAs (ceRNAs). Their closed-loop structure gives them high stability and their miRNA binding capacity seems much more powerful than that of any other ceRNAs, being named “super-sponges”. Today, few ceRNA prediction computational tools have been published but most of them do not consider ce-circRNAs [3,4]. Moreover, studies focusing on circRNA-miRNA-mRNA networks have not proposed an automated tool to search for sponge mechanisms involving circRNAs [5]. In this study, we present an interactive RShiny web application, called cirscan for circular RNA sponge candidates. cirscan automates the search for sponge mechanisms from input multi-level transcriptomic data (circRNA, miRNA, and mRNA) and represents the networks of interest as graphs, where nodes represent the different types of RNAs and edges the miRNA:target interactions. A major advance of the tool is that the user can query its own transcriptomic datasets from two specific conditions and apply filters, on both RNA expression levels and interaction scores, to prioritize the most likely sponge mechanisms. The identified mechanisms can be further investigated and may be a potential therapeutic target in human cancer. We applied cirscan on a previously published multi-level transcriptomic dataset from colorectal cancer (CRC). We identified 150 miRNAs potentially subjected to a sponge mechanism involving circRNAs and we retrieved a specific sponge mechanism previously described in the literature as involved in CRC carcinogenesis.

References

1. Anastasiadou E, Jacob LS, Slack FJ. Non-coding RNA networks in cancer. Nat Rev Cancer. janv 2018;18(1):5‑18.
2. Memczak S, Jens M, Elefsinioti A, Torti F, Krueger J, Rybak A, Maier L, Mackowiak SD, Gregersen LH, Munschauer M, Loewer A, Ziebold U, Landthaler M, Kocks C, le Noble F, Rajewsky N. Circular RNAs are a large class of animal RNAs with regulatory potency. Nature. 21 mars 2013;495(7441):333‑8.
3. Chiu HS, Llobet-Navas D, Yang X, Chung WJ, Ambesi-Impiombato A, Iyer A, Kim HR, Seviour EG, Luo Z, Sehgal V, Moss T, Lu Y, Ram P, Silva J, Mills GB, Califano A, Sumazin P. Cupid: simultaneous reconstruction of microRNA-target and ceRNA networks. Genome Res. févr 2015;25(2):257‑67.
4. Do D, Bozdag S. Cancerin: A computational pipeline to infer cancer-associated ceRNA interaction networks. PLOS Computational Biology. 16 juill 2018;14(7):e1006318.
5. Yi Y, Liu Y, Wu W, Wu K, Zhang W. Reconstruction and analysis of circRNA‑miRNA‑mRNA network in the pathology of cervical cancer. Oncol Rep. avr 2019;41(4):2209‑25.

Heterogeneous biological data integration with Semantic Web technologies using RDF or RDF-star formalisms generate topologically different graphs

Christophe Héligon[1], Olivier Dameron[2] and Jacques Pecreaux[1]
[1] CNRS, Univ Rennes, IGDR - UMR 6290, 2 avenue du Professeur Léon Bernard, F-35000, Rennes, France
[2] Univ Rennes, CNRS, Inria, IRISA - UMR 6074, 263 avenue du général Leclerc, F-35000, Rennes, France

Corresponding author: christophe.heligon@univ-rennes1.fr

Investigating the mechanisms of cell division and especially its fascinating robustness requires complex and time-consuming experimental approaches. It is, therefore, crucial to efficiently identify candidate genes involved in cell division. Using the asymmetric cell division of the nematode *Caenorhabditis elegans (C. elegans)* zygote, we set to integrate and process known biological data to prioritize sets of genes to be studied.
The roundworm *C. elegans* is a biological system in which detailed description and functional understanding require measuring multiple parameters (i.e. omics, phenotypes, interactions) at different scales (molecule, organelle, cell, organism, population) over time. Such biological data have accumulated and reached levels that are difficult to apprehend without appropriate informatics tools. While recently published tools use limited data types, we aim to take advantage of all kinds of available data and knowledge. Such heterogeneous multi-scale data integration is challenging in many aspects due firstly to different data formats and information content to be represented and secondly to the lack of standard integration tools to date.
We have identified the Semantic Web (SW) technologies as an appropriate general-purpose framework for data and metadata integration, processing and sharing compatible with the FAIR principles [1]. SW mediated data integration generates relational graphs using the Resource Description Framework (RDF) as a standard model for data description. RDF allows data merging even if the underlying schemas differ [2]. Integrating of our complementary heterogeneous datasets results in a complex data scheme (3921820 nodes and 10756313 relations). A significant part of this complexity results from the extensive use of properties describing the relations between entities, which require to introducing supplementary nodes to instantiate the relations. Recently, RDF-star has been proposed as a candidate extension of RDF that addresses this topic[3]. Therefore, we set to compare the original RDF representation of our datasets with its RDF-star counterpart.
The work presented here is a preliminary study to evaluate the consequences of choosing RDF or RDF-star for *C. elegans* data integration and processing. We take advantage of Wormbase.org, which provides a curated biomedical data repository for *C. elegans*. We compare RDF and RDF-star grammars for expressivity, general performance for data storage, SPARQL queries and graph topology.
We show here with the integration of 8 data types and 8 ontologies that classical RDF instantiations of relations and RDF-star formalism do not differ consistently for storage space, data endpoint upload time or query processing speed. Nevertheless, triple counts are smaller in RDF-star to represent complex data, and the graphs' topologies are very different. RDF's relation instantiation leads to the creation of generic nodes used as local intermediates between genes and biological data or metadata. In contrast, RDF-star implementation does not present such intermediates and genes are closer to biologically meaningful information. We still investigate whether discarding these intermediates improves the graphs' use by automated analysis tools as we envision using random walks and graph neural networks on the relational graphs in our prioritization application.

References

[1] Wim Hugo, Yann Le Franc, Gerard Coen, Jessica Parland-von Essen, and Luiz Bonino. D2.5 FAIR Semantics Recommendations Second Iteration, December 2020.

[2] Resource description framework (rdf). https://www.w3.org/RDF/. Accessed: 2022-05-15.

[3] Olaf Hartig. Foundations of rdf⋆ and sparql⋆ (an alternative approach to statement-level metadata in RDF). In Juan L. Reutter and Divesh Srivastava, editors, *Proceedings of the 11th Alberto Mendelzon International Workshop on Foundations of Data Management and the Web, Montevideo, Uruguay, June 7-9, 2017*, volume 1912 of *CEUR Workshop Proceedings*. CEUR-WS.org, 2017.

Revisiting iPPI-DB as a tool to navigate the pocketome protein-protein interactions

Fabien Mareuil[1], Alexandra Moine-Franel[2,4], Karen Druart[2], Bryan Brancotte[1], Rachel Torchet[1], Luis Checa Ruano[2,4] Hervé Ménager[1], Guillaume Bouvier[2], Vincent Mallet[2,3] and Olivier Sperandio[2]

[1] Institut Pasteur, Université Paris Cité, Bioinformatics and Biostatistics Hub F-75015 Paris, France
[2] Institut Pasteur, Université Paris Cité, Structural Bioinformatics Unit, Department of Structural Biology and Chemistry, CNRS UMR3528, Paris F-75015, France
[3] Center for Computational Biology, Mines ParisTech, Paris-Sciences-et-Lettres Research University, Paris 75272, France
[4] Collège Doctoral, Sorbonne Université, Paris F-75005, France

Corresponding author: fabien.mareuil@pasteur.fr

Our database iPPI-DB (inhibitors of Protein-Protein Interaction Database) [1] has historically been a repository of manually curated protein-protein interaction modulators from peer-reviewed articles or world patents and seminally focused on small molecules only. Hereafter, we describe a significant extension of iPPI-DB that allows users to investigate the wealth and diversity of putative PPI targets using a pocket-driven approach. In this revisited web application, a fully automated procedure has been developed to fetch all heterodimer complexes deposited in the PDB (Protein Data Bank), and map all binding pockets within the corresponding protein chains.

This new section contains the results of two different methods to detect and characterize pockets on a selected dataset from the PDB. The first method is based on InDeep [2] that relies on 3D fully convolutional neural networks to predict functional binding sites at the surface of proteins. These functional binding sites can take two forms, either an epitope binding site (location of a protein-protein interaction), or a druggable binding site (location for the binding of a future drug). The second method used, following on from the approach described in kuenemann et al. [3], uses VolSite [4] and RDKit [5] to detect the binding pockets and compute a set of descriptors calculated on the negative image of the binding site.

Besides the identification of these pockets, our approach evaluates their pair-wise similarities using a so-called Pocket Similarity Index (PSI) that relies on the pocket descriptors. This allows the inference of protein partners (epitopes or ligands) following the hypothesis that similar binding pockets might bind similar partners.

The web application allows to query the database from a pocket perspective (https://ippidb.pasteur.fr/targetcentric/): users can now query the dataset using a PDB code, and, from this query structure, access the most similar binding pockets available in the database, based on the aforementioned pocket properties.

References

[1] Rachel Torchet, Karen Druart, Luis Checa Ruano, Alexandra Moine-Franel, Hélène Borges, Olivia Doppelt-Azeroual, Bryan Brancotte, Fabien Mareuil, Michael Nilges, Hervé Ménager, et al. The ippi-db initiative: a community-centered database of protein–protein interaction modulators. *Bioinformatics*, 37(1):89–96, 2021.

[2] Vincent Mallet, Luis Checa Ruano, Alexandra Moine Franel, Michael Nilges, Karen Druart, Guillaume Bouvier, and Olivier Sperandio. Indeep: 3d fully convolutional neural networks to assist in silico drug design on protein–protein interactions. *Bioinformatics*, 38(5):1261–1268, 2022.

[3] Mélaine A Kuenemann, Céline M Labbé, Adrien H Cerdan, and Olivier Sperandio. Imbalance in chemical space: How to facilitate the identification of protein-protein interaction inhibitors. *Scientific reports*, 6(1):1–17, 2016.

[4] Jérémy Desaphy, Karima Azdimousa, Esther Kellenberger, and Didier Rognan. Comparison and druggability prediction of protein–ligand binding sites from pharmacophore-annotated cavity shapes, 2012.

[5] Greg Landrum, Paolo Tosco, Brian Kelley, Ric, sriniker, gedeck, Riccardo Vianello, NadineSchneider, Eisuke Kawashima, Andrew Dalke, Dan N, David Cosgrove, Brian Cole, Matt Swain, Samo Turk, AlexanderSavelyev, Gareth Jones, Alain Vaucher, Maciej Wójcikowski, Ichiru Take, Daniel Probst, Kazuya Ujihara, Vincent F. Scalfani, guillaume godin, Axel Pahl, Francois Berenger, JLVarjo, strets123, JP, and DoliathGavid. rdkit/rdkit: 2022_03_2 (q1 2022) release, April 2022.

APPINetwork : an R package for building and computational analysis of protein-protein interaction networks

Simon Gosset1[1], Annie Glatigny[2], Melina Gallopin[2], Marie-Hélène Mucchielli-Giorgi[1]

[1] Institut des sciences des plantes de Paris-Saclay,Batiment 630 rue Noetzlin, 91190, Gif sur Yvette, France

[2] Institut de Biologie Intégrative de la Cellule, 1 avenue de la terrasse, 91190, Gif sur Yvette, France

Corresponding Author: simon.gosset1@universite-paris-saclay.fr

Background

Protein-protein interaction (PPI) network analysis plays a major role in predicting the functionality of proteins and gives an insight into the functional relationships and evolutionary conservation of interactions among the proteins [1].

Hence, many tools have been developed to facilitate the construction and analysis of PPI networks. However, these tools presents some limits : Most of these tools do not allow the construction of PPI networks integrating both private and public data and do not use PPIs of degree 2. Moreover, they are often dedicated to a small number of model species and often do not provide a network analysis service.

To solve these deficiencies, we developed APPINetwork (Analysis of Protein-Protein Interaction Networks) a generic and user-friendly tool to build and analyse PPI network from different databases.

Methods

The APPINetwork package is an R package. It can be downloaded and installed from GitLab (https://forgemia.inra.fr/GNet/appinetwork). All functions are implemented in R except for the script that searches all proteins involved in a biological process (developed in C) and the scripts that format the BioGRID data file and generate the ID correspondence file (implemented in Python 3). The package can be deployed on Linux and macOS operating systems (OS). Deployment on Windows is possible but requires the prior installation of Rtools and Python 3. APPINetwork has a Graphical User Interface that facilitates its use by users that are not comfortable with terminals. This graphical interface is based on the widgets package. Graphical windows, buttons, and scroll bars allow the user to select or enter an organism name, select files and choose network parameters or methods dedicated to network analysis.

Results

APPINetwork allows PPI network building and analysis involving proteins from numerous biological process of numerous species or strains. It gives the possibility to the user to choose the public (experimental or predicted) PPI databases he/she wants to use to build the PPI network and to add his/her own PPI data.

It is a user-friendly package with many options that allow the building of a network adapted to the type of analysis to carry out: a network with or without genetic or predicted interactions in order to increase the number of PPIs which concern the studied biological process; a network of order one containing self-loops if the user wants to identify assembly intermediaries of a protein complex or a network of order two if he/she wants to identify sets of proteins involved in the same biological process. Other options of the interface allow to choose between the two types of analyse and their paramaters.

APPINetwork provides the PPI network as a flat file containing the list of PPIs with different information about the interaction and the proteins in interaction (PubMed IDs, experimental methods, all identifiers of involved proteins) that can be very useful for biologists. It provides also a text file containing all proteins of each cluster identified by Tfit and different files containing results of the hierarchical clustering modeling the assembly of a complex.

Finally, the package APPINetwork comes with a user guide and examples that facilitate its utilization.

References

[1] Paul Shannon. Cytoscape: A Software Environment for Integrated Models of Biomolecular Interaction Networks *Genome Res*, 13(11): 2498–2504, 2003.

PGxCorpus and PGxLOD: two shared resources for knowledge management in pharmacogenomics

Pierre Monnin[1,2], Joël Legrand[1,3] and Adrien Coulet[1,4,5]
[1] LORIA, Université de Lorraine, CNRS, Inria, Nancy, France
[2] Orange, Belfort, France
[3] CentraleSupélec, Metz, France
[4] Centre de Recherche des Cordeliers, Inserm, Univ. Paris Cité, Sorbonne Univ., Paris, France
[5] Inria Paris, Paris, France

Corresponding author: adrien.coulet@inria.fr

Pharmacogenomics (PGx) studies the impact of genetic factors on drug response phenotypes. Atomic knowledge units in PGx have the form of ternary relationships linking one or more drugs, one or more genetic factors, and one or more phenotypes. Such relationships state that a patient having the specified genetic factors and being treated with the specified drugs is likely to experience the given phenotypes. PGx knowledge is of particular interest for the development of precision medicine which aims at tailoring drug treatments to each patient to reduce adverse effects and maximize drug efficacy. However, PGx knowledge is scattered across many sources (*e.g.*, reference databases, the biomedical literature) and suffers from very heterogeneous levels of validation, *i.e.*, some PGx relationships are extensively studied and have been translated into clinical practice, but most are only observed on small-size cohorts or not reproduced yet and necessitate further investigation. Consequently, there is a strong interest in extracting and integrating knowledge units from these different sources into a unique place to provide a consolidated view of the state-of-the-art knowledge of this domain and drive to the validation, or moderation, of insufficiently validated knowledge units. To this aim, we created and share with the community two resources: PGxCorpus and PGxLOD.

PGxCorpus is a manually annotated corpus, designed for the automatic extraction of PGx relationships from text [1]. In specialized domains such as PGx, state-of-the-art approaches rely on supervised models trained or fine-tuned on previously annotated texts. PGxCorpus has been built by 11 annotators and consists of 945 sentences from PubMed abstracts annotated with *(i)* PGx entities of interest, *i.e.*, genetic factors (*e.g.*, genes, variants, haplotypes), drugs, and phenotypes, and *(ii)* relationships between these entities, associated with a type (*e.g.*, causes, decreases, transports) and an attribute (positive, hypothetical, or negative). It includes 2,875 relationships, each seen at least four times and in total by four different annotators. PGxCorpus is available at http://pgxcorpus.loria.fr.

PGxLOD is a knowledge graph that gathers 50,435 PGx relationships extracted from expert databases such as PharmGKB and from the literature [2]. It implements Semantic Web and FAIR best practices. Relationships of the literature are extracted with a model trained on PGxCorpus. Besides PGx relationships, PGxLOD includes knowledge about genetic factors, drugs, and phenotypes (*i.e.*, PGx key entities) imported from ClinVar, DisGeNET, DrugBank, SIDER, etc. to compose a graph of 88 million triples. We have paid particular attention at connecting PGx relationships that come from independent data sources but may be similar or equivalent with the development of both rule-based and machine learning matching approaches. PGxLOD is available at https://pgxlod.loria.fr.

These resources open perspectives with applications such as predicting pharmacogenes or mining molecular explanations of adverse drug reactions [3]. Additional analyses of PGxLOD offer the potential to guide PGx research by identifying knowledge that requires additional validation. Besides biomedical applications, PGxCorpus and PGxLOD offer challenging experimental settings to both NLP (discontinued entities, ternary relationships) and graph mining tasks (heterogeneity and arity of PGx relationships).

References

[1] J. Legrand et al. PGxCorpus, a manually annotated corpus for pharmacogenomics. *Sci. Data*, 7(1):3, 2020.

[2] P. Monnin et al. PGxO and PGxLOD: a reconciliation of pharmacogenomic knowledge of various provenances, enabling further comparison. *BMC Bioinformatics*, 20-S(4):139:1–139:16, 2019.

[3] E. Bresso et al. Investigating ADR mechanisms with explainable AI: a feasibility study with knowledge graph mining. *BMC Medical Informatics Decision Making*, 21(1):171, 2021.

Computational analysis of molecules, olfactory receptors, odors and their interactions.

Guillaume Ollitrault[1], Rayane Achebouche[1], Antoine Dreux[1], Karine Audouze[2], Anne Tromelin[3] and Olivier Taboureau[1]

[1] Inserm U1133, CNRS UMR 8251, Université de Paris, Paris, France
[2] T3S, Inserm UMR S-1124, Université de Paris, Paris, France
[3] Centre des Sciences du Goût et de l'Alimentation, AgroSup Dijon, CNRS, INRAE, Université Bourgogne Franche-Comté, Dijon, France

Corresponding Author: guillaume.ollitrault@inserm.fr

The sense of smell is a biological process involving volatile molecules that interact with proteins called olfactory receptors to transmit a nervous message that allows the recognition of a perceived odor. The integration of the relationships between molecules, olfactory receptors and odors is essential for a better understanding of their interactions. Based on 5907 odorant molecules, 98 olfactory receptors (human), 7029 odors and interactions between odorant molecules, odors and olfactory receptors, deep learning models have been developed, notably, Convolutional Neural Network (CNN) and Graphical Neural Network (GNN) and compared to Random Forest. The performance of such models is encouraging, with Precision/Recall Area Under Curve (PRC-AUC) values of 0.66 for odorant-odor GCN models and 0.91 for odorant-olfactory receptor GCN models. Such models should be able to predict the smell and olfactory receptors for a new molecule of interest. In addition, based on the encoding of the odorant molecule's structure, physicochemical features related to odors and/or olfactory receptors are proposed. Finally, the structural models of this set of olfactory receptors give us the possibility to apply a docking protocol and to suggest if a molecule can bind or not to an olfactory receptor. Considering all these approaches together, significant insights into olfaction are provided. We believe that our analysis is well suited to aid in the design of new odorant molecules and assist in fragrance research, sensory neuroscience and many other fields of research.

In-silico genome-wide detection of common fragile sites in human cells using BrdU-seq data.

Nathan ALARY[1], Stéphane KOUNDRIOUKOFF[2], Stefano GNAN[1], Michelle DEBATISSE[2] and Chun-Long CHEN[1]

[1] Institut Curie, Université PSL, Sorbonne Université, CNRS UMR 2344, F-75005 Paris, France

[2] Institut Gustave Roussy, CNRS UMR 8200, F-94805 Villejuif, France

Corresponding Author: chunlong.chen@curie.fr

Cancer cells display in general high levels of endogenous replication stress (RS), a major source of genome instability driving forces in the development of cancer [1]. In particular, RS specifically affects hard-to-replicate regions such as Common Fragile Site (CFSs), a type of evolutionary conserved sites that drive chromosome rearrangements during oncogenesis. CFSs display recurrent gaps or breaks on metaphase chromosomes of cells treated with aphidicolin, a DNA polymerases inhibitor that slows replisome progression [2]. CFSs are prevalent across late replicating regions of the genome and, importantly, frequently nest within large expressed genes, while all large expressed genes are not fragile [3]. Although complex, the relationships linking cell-type-specific transcription to fragility account for CFS tissue-specificity [4,5]. To date, since the detection of CFSs essentially relies on tedious cytogenetics analyses, they have been mapped mostly in human lymphoblasts and only a very limited number of other cell types, at a very low resolution (megabase scale). A simple high throughput mapping technique was therefore missing to identify CFSs instable in each type of cancer.

Our previous analysis of genome-wide replication timing in human lymphoblasts has shown that aphidicolin induces delayed and/or under-replication of specific regions nested in late-replicating large genes (Significantly Delayed Regions; SDRs) [3]. Strikingly, we found that SDRs precisely co-map with CFSs previously mapped in these cells. Based on this finding, we have now developed a simple method and corresponding bioinformatic tool allowing genome-wide mapping of CFSs. In this method, newly synthesized DNA is classically recovered from cells pulse-labeled with BrdU. Labeled DNA from untreated cells and aphidicolin-treated cells is immunoprecipitated and sequenced (BrdU-seq) as previously described. Importantly, we demonstrated that SDRs can be identified without sorting cells at different steps of the S phase by detecting significantly and differentially depleted regions of the genome, which considerably simplifies the CFS mapping process.

Genome-wide exploration of CFSs in different cell types may reveal uncharacterized cancer-associated CFSs genes, providing new insights into the mechanisms underlying human diseases. Further elucidation of the clinical significance and biological functions of these genes may be exploited for cancer biomarkers and therapeutic benefits.

References

[1] Magdalou I, Lopez BS, Pasero P, and Lambert SA (2014). The causes of replication stress and their consequences on genome stability and cell fate. *Semin. Cell Dev. Biol* 30, 154–164.

[2] Durkin SG, Glover TW. Chromosome fragile sites. *Annu. Rev. Genet.* 2007;41:169–192.

[3] Olivier Brison, Sami El-Hilali, *et al*.. Transcription-mediated organization of the replication initiation program across large genes sets common fragile sites genome-wide. *Nature Communication*, 2019, 10, pp.5693.

[4] Le Tallec B, *et al*. Common fragile site profiling in epithelial and erythroid cells reveals that most recurrent cancer deletions lie in fragile sites hosting large genes. *Cell Rep*. 2013;4:420–428.

[5] Wilson TE, et al. Large transcription units unify copy number variants and common fragile sites arising under replication stress. *Genome Res*. 2015;25:189–200.

Evaluation of word-based alignment-free methods for yeast genome comparison and taxonomy

Emmanuel BUSE FALAY[1], Jean-Luc LEGRAS[1] and Hugo DEVILLERS[1]

[1]SPO, Univ Montpellier, INRAE, Institut Agro, 2 Place Pierre Viala, 34060, Montpellier, France

Corresponding Authors: emmanuelbuse@outlook.fr , hugo.devillers@inrae.fr

Background

Sequence comparison is one of the fundamental tasks to study the evolutionary relationships between organisms or groups of organisms. Methods based on pairwise or multiple alignment of molecular sequences are traditionally used to address this problem. In recent years, the availability of large amounts of genomic data produced by new sequencing technologies offers the opportunity to compare large sets of whole genome sequences. In this context, alignment-based methods can fail to achieve such comparison within a reasonable computation time. The alignment-free (AF) methods [1] are good alternatives to study evolutionary relationships between organisms. Several AF sequence approaches, including those based on the number of words or k-mers, have been successfully employed in different studies, showing their interest in particular for whole-genome comparison, taxonomy and phylogenomics in a correct computation time. Moreover, contrary to alignment-based approaches, AF approaches have the advantage of taking into account all the genomic information and do not require the prior identification of alignable homologous segments from the genomes to be compared.

AF approaches have proven their worth in numerous comparative and analytical applications on sequences from a wide range of organisms, from small virus genomes to large Eukaryotic genomes or even on metagenomic datasets. More specifically, while AF methods have been successfully applied on taxonomic/phylogenomic approaches on viruses and Prokaryotes, such applications on Eukaryotic genomes remain rarely used. In this context, here we present an evaluation of the interest and performances of AF methods to analyze genomes of ascomycetes yeasts. Our aims were to evaluate the reliability of AF methods to classify yeast strains from the same species, to test their ability to work directly with sequencing data and to identify the best suited AF tool as well as its optimal parameter values (e.g., k-mer length).

Approach

Five tools (SANS Serif, Mash, CAFE, kSNP and AAF) implementing different k-mer based alignment-free methods rooted on different algorithms were selected and tested on three types of data (raw sequencing data, cleaned sequencing data and assembled genomes) from 68 strains of the yeast species *Torulaspora delbrueckii*.

To evaluate the reliability of AF methods predictions, all produced results were confronted to a reference tree based on a classical SNP phylogenetic approach [2].

Results

The alignment-free tools selected in this study demonstrated variable performances to classify yeast strains from *T.delbruekii*. However, some of them, with appropriate parameter values succeeded to produce very accurate classifications in comparison to the reference tree, even when considering raw sequencing data as input. In the present work, the different results we obtained are presented and discussed.

References

1. Zielezinski A, Vinga S, Almeida J, Karlowski WM. Alignment-free sequence comparison: benefits, applications, and tools. Genome Biol. 2017;18:186.

2. Silva M, Pontes A, Franco-Duarte R, Soares P, Sampaio JP, Sousa M, et al. A glimpse at an early stage of microbe domestication revealed in the variable genome of Torulaspora delbrueckii, an emergent industrial yeast. preprint. Preprints; 2021.

MPS-Sampling : a novel method allowing the reliable selection of representative genomes to infer large-scale phylogenies

Rémi-Vinh COUDERT[1,2], Frédéric JAUFFRIT[2], Jean-Philippe CHARRIER[2], Jean-Pierre FLANDROIS[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

[2] Microbiology Research, bioMérieux SA, Marcy-l'Étoile

Corresponding Author: remi.coudert@univ-lyon1.fr

Molecular phylogenetics has spread to all areas of life sciences: evolutionary biology, ecology, cell biology, biochemistry, microbiology since it allows observations or results to be placed in a more general framework. In that context, inferring reliable phylogenies of organisms is a major challenge, especially for prokaryotes [1]. Most reference phylogenies are inferred using sets of conserved single copy universal markers using either super-tree or super-matrix approaches [2]. Among them ribosomal proteins are routinely used [3], because they are universal, very conserved, less prone to horizontal gene transfers compared to other markers [4].

The recent explosion of genome sequencing projects opens up unique opportunities by providing a wealth of data and ever-increasing access to the genetic information. The disadvantage is that it also makes the assembly of the data sets more complex especially if we planned to use the most efficient reconstruction methods (maximum likelihood, Bayesian methods) and the most realistic evolutionary models (CAT, C60...). An appropriate selection of a subset of representative genomes, from all available genomes, can help overcome this problem.

Most of the time, the selection of representative genomes is based on academic knowledge, such as taxonomy with some subjectivity. A classical approach is to select one representative strain per species. The main limitations of these approaches are that (i) the criteria for defining species can vary, (ii) misclassifications are frequent, (iii) the number of species (~16,000) is still too large and (iv) genomes may not be identified to a species.

Here, we present Multiple Protein Similarity Sampling **(MPS-Sampling)**, an automated and *ex nihilo* method of representative genome selection that do not rely on environmental, academic, historical, or cultural criteria. First of all, the sequences of each protein are clustered individually, leading to several divisions, one per protein. Then, the clusterings of each protein are harmonized together at the genome level. Finally, a representative genome is selected on the basis of centrality and quality criteria.

To illustrate this method, MPS-Sampling was used to infer reference phylogenies of *Bacteria* and *Archaea* using ribosomal proteins contained in RiboDB [5]. RiboDB is a dedicated database gathering all ribosomal proteins families, involving 182,496 complete archaeal and bacterial genomes in the last release of June 2022. More precisely, we selected a representative subset of 3,451 *Bacteria* genomes (2%) in 2 hours to infer a reliable phylogeny representing 158,000 genomes.

References

[1] S. Gribaldo, C. Brochier, Phylogeny of prokaryotes: does it exist and why should we care?, Research in microbiology, 160 (2009) 513-521.

[2] F. Delsuc, H. Brinkmann, H. Philippe, Phylogenomics and the reconstruction of the tree of life, Nat Rev Genet, 6 (2005) 361-375.

[3] T.A. Williams, C.J. Cox, P.G. Foster, G.J. Szollosi, T.M. Embley, Phylogenomics provides robust support for a two-domains tree of life, Nat Ecol Evol, 4 (2020) 138-147.

[4] C. Brochier, E. Bapteste, D. Moreira, H. Philippe, Eubacterial phylogeny based on translational apparatus proteins, Trends in genetics : TIG, 18 (2002) 1-5.

[5] F. Jauffrit, S. Penel, S. Delmotte, C. Rey, D.M. de Vienne, M. Gouy, J.P. Charrier, J.P. Flandrois, C. Brochier-Armanet, RiboDB Database: A Comprehensive Resource for Prokaryotic Systematics, Molecular biology and evolution, 33 (2016) 2170-2172.

Alternative splicing modulates the number and composition of similar exonic regions

Antoine Szatkownik[1,2], Hugues Richard[1,2] and Elodie Laine[1]

[1] Sorbonne Universite´, CNRS, IBPS, Laboratoire de Biologie Computationnelle et Quantitative (LCQB), 75005 Paris, France

[2] Bioinformatics Unit (MF1), Department for Methods Development and Research Infrastructure, Robert Koch Institute, 13353 Berlin, Germany

Corresponding Author: RichardH@rki.de, elodie.laine@sorbonne-universite.fr

The duplication of genetic material and the production of different transcript isoforms from the same gene locus are two major mechanisms for generating protein diversity in multicellular eukaryotes. At the human protein-coding genome scale, we recently identified a few thousand genes where the alternative splicing (AS) of pre-mRNA transcripts modulates the number and composition of similar exonic regions [1]. The corresponding proteins are involved in muscle contraction and neural intercellular communication. This alternative usage of "pseudo-repeats" is evolutionary conserved and influences partner binding affinity, specificity, and stoichiometry.

Here, we report on a method for automatically detecting alternatively spliced pseudo-repeats from an ensemble of transcripts observed in a set of genes/species. It builds on a graph-based data structure we introduced in [1] where the nodes, called “s-exons”, represent minimal transcripts’ building blocks defined across several species. Formally, a s-exon is a multiple sequence alignment of sub-exons or sub-exon parts coming from different genes/species. The present work expands on the notion of s-exon by defining evolutionary meaningful AS-aware pseudo-repeat units (ASRUs). Each ASRU is a collection of s-exons that share some similarity and are alternatively used in the observed transcripts. We detected 1070 evolutionary conserved ASRUs coming from 717 genes over the whole human proteome –over one order of magnitude more than previously reported [2]. We observed a wide spectrum of scenarios ranging from a single ASRU comprising 2 instances, which is the case for most genes, to extreme cases like Nebulin with a single ASRU containing over 80 instances, or Myosin heavy chain with 36 ASRUs, each comprising two instances. By performing a contrastive analysis of ASRUs conservation across protein families in a species, as well as across species, we can put in relation sequence conservations with the properties of interaction specificity-determining sites. Those analyses can shed light on the balance between the repetition of functional elements in a genome and the amount of selective pressure they are subject to.

References

1. Diego Javier Zea, Sofya Laskina, Alexis Baudin, Hugues Richard and Elodie Laine. Assessing conservation of alternative splicing with evolutionary splicing graphs. *Genome Res*. 2021 31:1462-1473, doi :10.1101/gr.274696.120.

2. Federico Abascal, Michael L. Tress, and Alfonso Valencia. The Evolutionary Fate of Alternatively Spliced Homologous Exons after Gene Duplication. Genome Biol Evol 7: 1392–1403. doi:10.1093/gbe/evv076

MicroScope: a web platform for microbial genome annotation through pangenomic and metabolic analysis

Alexandra Calteau[1], Mathieu Dubois[1], Jérôme Arnoux[1], Lucie Coffion[1], Stéphanie Fouteau[1], Aurélie Lajus[1], Clovis Norroy[1], David Roche[1], Zoé Rouy[1], Mark Stam[1], Hannah Tomelka[1], Claudine Médigue[1], David Vallenet[1]

[1] LABGeM, Génomique Métabolique, CEA, Genoscope, Institut François Jacob, Université d'Évry, Université Paris-Saclay, CNRS

Corresponding Author: acalteau@genoscope.cns.fr

1. Introduction

Large-scale genome sequencing projects are producing a vast amount of new information that is completely transforming our understanding of thousands of microbial species. However, despite the development of powerful bioinformatics approaches, full interpretation of the content of these genomes remains a difficult task for microbiologists.

To address this challenge, we develop the MicroScope platform, an integrated Web platform for management, annotation, comparative analysis and visualization of microbial genomes (https://mage.genoscope.cns.fr/microscope) [1]. The platform enables collaborative work in a rich comparative genomic context and improves community-based curation efforts.

2. Methods

Launched in 2005, the platform has been under continuous development. MicroScope provides analyses for complete and ongoing genome projects together with metabolic network reconstruction and transcriptomic experiments allowing users to improve the understanding of gene functions.

Besides automatic functional annotations, several tools allow analyzing a wide range of biological systems (antibiotic resistance, secondary metabolites, secretions systems, defense systems,...). The platform also has extensive functionalities to explore and compare metabolic pathways. Finally, recent functionalities allow users to perform comparative pangenomics on hundreds of genomes of the same species and to explore their content in regions of genomic plasticity.

3. Results

MicroScope platform is widely used by microbiologists from academia and industry all around the world for collaborative studies and expert annotation. To date, MicroScope contains data for >16,000 microbial genomes, part of which are manually curated and maintained by microbiologists (>6,100 user accounts in May 2022). The platform is also a useful resource for academic training.

Acknowledgements

The France Génomique and French Bioinformatics Institute national infrastructures (funded as part of Investissement d'Avenir program managed by Agence Nationale pour la Recherche, contracts ANR-10-INBS-09 and ANR-11-INBS-0013) are acknowledged for their support of the MicroScope annotation platform.

References

1. David Vallenet, Alexandra Calteau, Mathieu Dubois, Paul Amours, Adelme Bazin, Mylène Beuvin, Laura Burlot, Xavier Bussell, Stéphanie Fouteau, Guillaume Gautreau, Aurélie Lajus, Jordan Langlois, Rémi Planel, David Roche, Johan Rollin, Zoe Rouy, Valentin Sabatet and Claudine Médigue. MicroScope: an integrated platform for the annotation and exploration of microbial gene functions through genomic, pangenomic and metabolic comparative analysis. *Nucleic Acids Research*, 48:D579–D589, 2020.

Comparative study of protein aggregation propensity and mutation tolerance between naked mole-rat and mouse

Savandara Besse[1], Raphaël Poujol[2] and Julie Hussin[3]

[1] Adrian Serohijos' Evolutionary Biophysics Lab, Université de Montréal, Département de Biochimie et Médecine Moléculaire, H3T 1J4, Montréal, Canada

[2] Julie Hussin's Computational Biomedecine Lab, Institut de Cardiologie de Montréal, H1T 1C8, Montréal, Canada

[3] Julie Hussin's Computational Biomedecine Lab, Institut de Cardiologie de Montréal, H1T 1C8, Montréal, Canada

Corresponding Author: savandara.besse@umontreal.ca

The molecular mechanisms of aging and life expectancy have been studied in model organisms with short lifespans. However, long-lived species may provide insights into successful strategies of healthy aging, potentially opening the door for novel therapeutic interventions in age-related diseases. Notably, naked mole-rats, the longest-lived rodent, present attenuated aging phenotypes in comparison to mice. Their resistance toward oxidative stress has been proposed as one hallmark of their healthy aging, suggesting their ability to maintain cell homeostasis, and specifically their protein homeostasis. To identify the general principles behind their protein homeostasis robustness, we compared the aggregation propensity and mutation tolerance of naked mole-rat and mouse orthologous proteins. Our analysis showed no proteome-wide differential effects in aggregation propensity and mutation tolerance between these species, but several subsets of proteins with a significant difference in aggregation propensity. We found an enrichment of proteins with higher aggregation propensity in naked mole-rat involved the inflammasome complex, and in nucleic acid binding. On the other hand, proteins with lower aggregation propensity in naked mole-rat have a significantly higher mutation tolerance compared to the rest of the proteins. Among them, we identified proteins known to be associated with neurodegenerative and age-related diseases. These findings highlight the intriguing hypothesis about the capacity of the naked mole-rat proteome to delay aging through its proteomic intrinsic architecture.

Acknowledgements

This work was supported by funds from the Department of Biochemistry and Molecular Medicine of Université de Montréal and through the access to computational resources provided by Calcul Québec to JGH. We thank Sebastian Pechmann for his supervision on the preliminary analyses. We are grateful to Adrian Serohijos for his mentoring support and the fruitful discussions throughout the project. Finally, we thank all the members of the Hussin lab for their constructive comments and feedback on the figures for the manuscript. JGH is a Fonds de la Recherche du Québec en Santé (FRQS) Junior 1 Scholar, funded by the Institute for Data Valorization (IVADO).

Yeast recombination specificity impact on demography inference

Louis OLLIVIER[1], Flora JAY[1] and Fanny POUYET[1]

[1] LISN, Rue Raimond Castaing, 91190, Gif-Sur-Yvette, France

Corresponding Author: louis.ollivier@etu.univ-rouen.fr

The canonical life cycle of Saccharomyces cerevisiae consists of an alternation between mitosis (clonal reproduction) and meiosis events (outcrossing and intratetrad mating). In this life cycle, meiosis events and then recombination events are known to be rare. This raises questions because it is known that genetic recombination favors genetic diversity.

However, recent studies [1] start to show that heterozygosity in yeast would be greater than previously thought, which implies that recombination events might be more frequent than anticipated. Indeed, these events are estimated at frequencies ranging from 1 per 50000 generations to 1 per 1000 generations (the difficulty to estimate these frequencies makes it hard to give a more accurate range).

Using SLiM [2] and dadi [3], we investigate if we can correctly infer demographic parameters given the frequency of recombination events. We test multiple demographic scenarios from simple ones (exponential growth or bottleneck in a single population, …) to more complex ones (multiple populations with migration [4], …).

For instance, we observe interesting results such as an excess of polymorphic sites at frequency 0.5 for simulations of populations with bottleneck demography and recombination events every 1000 generations.

References

1. Fischer, G, Liti, G, Llorente, B. The budding yeast life cycle: More complex than anticipated? *Yeast*; 38: 5– 11, 2021
2 . Haller, B.C., & Messer, P.W. (2019). SLiM 3: Forward genetic simulations beyond the Wright–Fisher model. *Molecular Biology and Evolution 36*(3), 632–637
3. RN Gutenkunst, RD Hernandez, SH Williamson, CD Bustamante "Inferring the joint demographic history of multiple populations from multidimensional SNP data" PLoS Genetics 5:e1000695 (2009)
4. Duan, SF., Han, PJ., Wang, QM. *et al.* The origin and adaptive evolution of domesticated populations of yeast from Far East Asia. *Nat Commun* **9,** 2690 (2018).

Wood-decomposing fungi through the lens of genomes comparison

Annie LEBRETON[1], Otto MIETTINEN[2], Elodie DRULA[3], Marie-Noelle ROSSO[3], Sajeet HARIDAS[4], Jasmyn PANGILINAN[4], Anna LIPZEN[4], Marc BUÉE[1], Annegret KOHLER[1], Kerrie BARRY[4], Igor V. GRIGORIEV[4,5], Håvard KAUSERUD[6], Francis MARTIN[1] and Sundy MAURICE[1,6]

[1] Université de Lorraine, INRAE, UMR1136, InteractionsArbres/Microorganismes, 54280, Champenoux, France

[2] Finnish Museum of Natural History, P.O. Box 7, FI-00014 University of Helsinki, Finland

[3] INRAE, Aix Marseille Univ, UMR1163, Biodiversité et Biotechnologie Fongiques, 13009, Marseille, France

[4] US Department of Energy Joint Genome Institute, Lawrence Berkeley National Laboratory, Berkeley, 94720, California, USA

[5] Department of Plant and Microbial Biology, University of California Berkeley, 94720, Berkeley CA, USA

[6] Section for Genetics and Evolutionary Biology, Department of Biosciences, University of Oslo, Blindernveien 31, 0316, Oslo, Norway

Corresponding Author: annie.lebreton@inrae.fr

Fungi play key roles in the nutrient cycles of boreal, temperate and tropical forests. In these biomes, many wood-decay fungi have a narrow host range, colonizing either coniferous or deciduous host trees. Some fungal species have the capacity to colonize and decompose both types, hence increasing their chances to find a suitable substrate and may contribute to its fitness. In species with such wide ecological niches, specializing and adaptation towards certain substrates might develop over time, likely dependent on the availability of host trees in a biome. To identify potential genomic markers associated to host specialization, we analysed a set of 132 fungal genomes covering 14 orders in the Agaricomycetes class with a focus on the Polyporales and Agaricales (67 and 22 species respectively). Phylogenomic analyses identified the position of the newly sequenced species. Complementary annotations of genes potentially associated to wood-decay mechanisms (e.g. CAZymes) and interaction with the host tree (e.g. secretome and GPCR) were performed. Their comparisons highlighted clear differences within and among the different Agaricomycetes orders, such as the differences in plant degrading enzyme content between two distinct types of wood decay: the white rot and brown rot species. Gene families associated to niche breadth evolution such as those specifically enriched or depleted in coniferous and deciduous species were also identified. These results contribute in building a framework for understanding the genetic determinants of the evolution of fungal decay mechanisms and, potentially, predicting the responses of fungal communities to substrate and habitat pressures in a forest management and climate change context.

Caractérisation du contenu en gènes de l'espèce bactérienne Coxiella *burnetii* issue de différentes lignées en Europe

Aminah Keliet[1], Karim Sidi-Boumedine[2], Elsa Jourdain[1], Richard Thiéry[2], Elodie Rousset[2], Xavier Bailly[1]

1 Université Clermont Auvergne, INRAE, VetAgro Sup, UMR EPIA Epidémiologie des maladies animales et zoonotiques, 63122, Saint-Genès-Champanelle, France
2 Anses (French Agency for Food, Environmental, and Occupational Health and Safety), Laboratory of Sophia Antipolis, Animal Q Fever Unit, Sophia Antipolis, France.

Corresponding Author: aminah.keliet@inrae.fr

Coxiella *burnetii* est la bactérie responsable de la fièvre Q, une zoonose répandue et pouvant infecter un large spectre d'espèces hôtes. L'infection est souvent asymptomatique mais chez les humains, elle peut entraîner divers troubles cliniques problématiques et parfois persistants. Elle est transmise par voie aérienne à partir de l'environnement contaminé par les réservoirs animaux, dont les principaux sont les ruminants domestiques. En Europe, on observe une spécificité d'hôte, qui se traduit par l'association de lignées de C. *burnetii* avec différents ruminants domestiques [1].
Le but de l'étude est donc d'étudier la dynamique des génomes de *C. burnetii* afin de contribuer aux connaissances sur l'émergence de lignées caractérisées par différentes spécificités d'hôtes.

Nous avons effectué ces analyses à partir de souches de référence et de souches provenant de prélèvements dans des fermes. Nous avons réalisé i) une identification des gènes homologues et une sélection des gènes communs pour obtenir une phylogénie. Pour effectuer l'identification des gènes homologues, nous avons étudié le pangénome de C. *burnetii* en comparant les pipelines BPGA et Panaroo ; ii) une étude de la distribution des gènes et fonctions accessoires par analyse factorielle des correspondance (AC) à partir de la matrice de présence/absence de gènes ; iii) une étude de la dynamique de distribution de ces gènes le long de la phylogénie, avec l'outil CAFE3.

L'identification des gènes homologues obtenue avec Panaroo nous semble la plus cohérente, l'utilisation de BPGA pointant un nombre important de gènes uniques en proportion très variable en fonction des génomes. L'arbre phylogénétique obtenu montre la présence de trois grandes lignées au sein des souches de *C. burnetii* étudiées, nommées A, B et C [1]. L'analyse de la matrice de présence absence de gènes par AC illustre la cohérence de la distribution des gènes dans les lignées. La distribution des gènes dans le groupe C est particulièrement homogène, le nombre de gènes accessoires spécifiques est limité et beaucoup sont présents dans la plupart des souches. La distribution des gènes semble plus hétérogène dans les groupes A et B. Le groupe A inclut une grande diversité de gènes dont une large proportion de gènes spécifiques. Les gènes accessoires du groupe B sont particulièrement différenciés par rapport à ceux du groupe C. Au cours de la diversification de *C. burnetii*, les analyses effectuées pointent différents évènements de perte massive de gènes, notamment au niveau des ancêtres communs de la lignée C et des lignées A et C.

L'analyse des gènes effecteurs et de gènes de virulence identifiés par ailleurs montre que ces gènes, potentiellement impliqués dans l'interaction avec l'hôte, peuvent présenter un polymorphisme de présence/absence de l'ordre de 5% et 0,8%, respectivement. De futures études sont nécessaires pour identifier si ces polymorphismes sont impliqués dans les différences de spécificité d'hôte entre lignées ou si d'autres déterminants doivent être recherchés.

References

1. Joulié A, Karim Sidi-Boumedine, Xavier Bailly, Patrick Gasqui, Séverine Barry, Lydia Jaffrelo, Charles Poncet, David Abrial, Elise Yang, Animal diagnostic laboratories consortium, Leblond A, Rousset E, Jourdain E. Molecular epidemiology of Coxiella burnetii in French livestock reveals the existence of three main genotype clusters and suggests species-specific associations as well as regional stability. Infect Genet Evol. 2017 Mar;48:142-149. doi: 10.1016/j.meegid.2016.12.015. Epub 2016 Dec 20. PMID: 28007602.

The genomic basis of the Streptococcus thermophilus health-promoting properties

Emeline ROUX[2,3,4], Aurélie NICOLAS[1], Florence VALENCE[1], Grégoire SIEKANIEC[1,3], Victoria CHUAT[1], Jacques NICOLAS[3], Yves LE LOIR[1] and Eric GUÉDON[1]

[1]INRAE, Institut Agro, STLO, Rennes, France

[2] CALBINOTOX, Université de Lorraine, F-54000, Nancy, France

[3] Univ Rennes, Inria, CNRS, IRISA, Rennes F-35000, France

[4] Institut NuMeCan, INRAE, INSERM, Univ Rennes, Saint-Gilles, France

Corresponding Author: eric.guedon@inrae.fr

1. Introduction

Streptococcus thermophilus is a Gram-positive bacterium widely used as starter in the dairy industry as well as in many traditional fermented products. In addition to its technological importance, it has also gained interest in recent years as beneficial bacterium due to human health-promoting functionalities. The objective of this study was to inventory the main health-promoting properties of *S. thermophilus* and to study their intra-species diversity at the genomic and genetic level within a collection of 79 representative strains.

2. Methods

FastANI software was used to compute pair-wise ANI (Average Nucleotide Identity) values to confirm that the 79 selected genomes belong to the same species. MicroScope Pan-genome analysis tool and eggnog-mapper were used to analyse core and accessory genome. Text mining was used to highlight pseudogene annotations. AntiSMASH and BAGEL4 were used to search bacteriocins sequences and ABRicate to search antibiotic resistance genes. R and Python libraries were used to create heatmaps.

3. Results

In this study various health-related functions were analyzed at the genome level from 79 genome sequences of strains isolated over a long time period from diverse products and different geographic locations. While some functions are widely conserved among isolates (e.g., degradation of lactose, folate production) suggesting their central physiological and ecological role for the species, others including the tagatose-6-phosphate pathway involved in the catabolism of galactose, and the production of bioactive peptides and gamma-aminobutyric acid are strain-specific. Most of these strain-specific health-promoting properties seems to have been acquired via horizontal gene transfer events. The genetic basis for the phenotypic diversity between strains for some health-related traits have also been investigated. For instance, substitutions in the *galK* promoter region correlate with the ability of some strains to catabolize galactose via the Leloir pathway. Finally, the low occurrence in *S. thermophilus* genomes of genes coding for biogenic amine production and antibiotic resistance is also a contributing factor to its safety status.

4. Conclusions

The natural intra-species diversity of *S. thermophilus*, therefore, represents an interesting source for innovation in the field of fermented products enriched for healthy components. A better knowledge of the health-promoting properties and their genomic and genetic diversity within the species may facilitate the selection and application of strains for specific biotechnological and human health-promoting purpose. Moreover, by pointing out that a substantial part of its functional potential still defies us, our work opens the way to uncover additional health-related functions through the intra-species diversity exploration of *S. thermophilus* by comparative genomics approaches.

Evolution is not uniform along coding sequences

Raphaël BRICOUT[1], Dominique WEIL[2], David STROEBEL[1], Auguste GENOVESIO[1]*, Hugues ROEST CROLLIUS[1]*

[1] Institut de Biologie de l'ENS (IBENS), Département de biologie, École normale supérieure, CNRS, INSERM, Université PSL ; 46 rue d'Ulm, 75005 Paris, France

[2] Institut de Biologie Paris Seine (IBPS)

* Corresponding authors: auguste.genovesio@ens.psl.eu, hrc@bio.ens.psl.eu

Abstract

Rates of evolution vary between protein-coding gene families, but also within gene families and even along the length of coding sequences, because the functional and structural roles of amino acids are different [1]. However the impact of the amino acid positions inside the sequence on the rate of molecular evolution has not been yet investigated. We estimate average evolutionary rates for position-specific codons in primate coding sequences, and discovered a strong excess of non-synonymous substitution at sequences termini compared to the center. This bias was also observed in plant gene families, suggesting that it is universal in eukaryote genes. Control experiments excluded the possibility that annotation errors, alignment artifacts and compositional bias would cause the observed profile. We further show that the distribution of functional domains [2] and of solvent-accessible residues in proteins [3] readily explain how functional constrains are weaker at their termini, leading to the observed excess of amino-acid substitutions. Finally, we show that methods inferring sites under positive selection are strongly biased towards protein termini, suggesting that they may confound positive selection with weak negative selection. These results suggest that accounting for positional information should improve evolutionary models.

Acknowledgements

We thank P. Vincens and the informatics service at IBENS for support, and Alexandra Louis, Guillaume Louvel, François Giudicelli and Nicolas Lartillot for helpful discussions.

References

[1] J. Echave, S. J. Spielman, C. O. Wilke, *Nat Rev Genet*. **17**, 109–121 (2016).

[2] Y. Wang, H. Zhang, H. Zhong, Z. Xue, *Computational and Structural Biotechnology Journal*. **19**, 1145–1153 (2021).

[3] J. Jumper *et al.*, *Nature*. **596**, 583–589 (2021).

Most of the genetic diversity of the *Wolbachia* infecting *Culex pipiens* lies in the prophage regions

Camille MESTRE*, Alice NAMIAS*, Mathieu SICARD*, Mylène WEILL*
* ISEM, Université de Montpellier, CNRS, IRD, Place Eugène Bataillon, 34095, Montpellier, France

Corresponding Autor: camille.mestre@umontpellier.fr

1. Introduction

Wolbachia are maternally-transmitted endosymbiotic bacteria which infect up to 60% of insect species [1,2]. In some of its hosts, including the mosquito *Culex pipiens*, *Wolbachia* induces cytoplasmic incompatibility (CI), a conditional sterility by which crosses between two hosts infected by so-called incompatible *Wolbachia* will be sterile. The description, based on 7 polymorphic genes, of 5 phylogenetic groups of *w*Pip (the *Wolbachia* infecting *Cx. pipiens*) shed light on the highly complex CI patterns described in *Cx. pipiens*: crosses between mosquitoes infected by *w*Pip of the same group are compatible, while inter-group crosses have unpredictable outcomes [3,4].

2. Data set

We had access to the data of the 1000 *Culex pipiens* genomes project led by Lindy McBride and Yuki Haba which include genomic material of mosquitos and their endosymbionts. We thus used approximately 800 samples of paired-end DNA Illumina short reads from all over the world to build *w*Pip phylogenies, based on the full genome (1,482,455 bp), rather than 7 polymorphic genes. We also looked at groups' geographic distributions.

3. Results

An individual was analyzed if more than 30% of *Wolbachia* positions were covered, aligning the reads on *w*Pip reference genome Pel [5]. We identified the *Wolbachia* group infecting each individual by identifying the PK1 allele, using BLAST command line and the assembler MEGAHIT. This gene code for ankyrin domain protein and its variations was found to correspond to the *Wolbachia* group [4]. We assembled the *w*Pip genome present in each individual, and performed a PCA on full genomes, keeping only biallelic SNPs. We found that individuals still cluster into the 5 expected groups (which corresponded to PK1 groups). Looking at highly contributing SNPs on the first three PCA axes, we found that all those SNPs were located in the previously annotated prophage regions [3]. We removed these regions and found, with a new PCA, that individuals still clustered by genetic groups. Yet, we found that highly contributing SNPs of this PCA, are still in annotated phage genes, located outside of prophage regions. Finally, we showed that coexisting *w*Pip from distinct groups differed more that *w*Pip from the same group from distinct places, and no isolation-by-distance was described within *w*Pip groups, suggesting a recent radiation.

4. Conclusion

Using all wPip genome sequences we confirmed the recent radiation of the five wPip phylogenetic groups. Previous studies described congruence between *w*Pip phylogenetic trees and mitochondrial trees. We will examine this on a much broader dataset, to identify cases of paternal or horizontal *w*Pip transmission.

Acknowledgements

We thank Yuki Haba and Lindy McBride for the unlimited access to the data. We also want to thank Michael Turelli, Will Conner, Khalid Belkhir and Rémy Dernat for their help at various stages of this project.

References

[1] Zug, R., & Hammerstein, P. (2012). Still a Host of Hosts for Wolbachia : Analysis of Recent Data Suggests That 40% of Terrestrial Arthropod Species Are Infected. *PLoS ONE*, *7*(6), e38544.
[2] Hilgenboecker, K., Hammerstein, P., Schlattmann, P., Telschow, A., & Werren, J. H. (2008). How many species are infected with Wolbachia? – A statistical analysis of current data: Wolbachia infection rates. *FEMS Microbiology Letters*, *281*(2), 215-220.
[3] Atyame, C. M., Delsuc, F., Pasteur, N., Weill, M., & Duron, O. (2011). Diversification of Wolbachia Endosymbiont in the Culex pipiens Mosquito. *Molecular Biology and Evolution*, *28*(10), 2761-2772.
[4] Dumas, E., Atyame, C. M., Milesi, P., Fonseca, D. M., Shaikevich, E. V., Unal, S., Makoundou, P., Weill, M., & Duron, O. (2013). Population structure of Wolbachia and cytoplasmic introgression in a complex of mosquito species. *BMC Evolutionary Biology*, *13*(1), 181.
[5] Klasson, L., Walker, T., Sebaihia, M., Sanders, M. J., Quail, M. A., Lord, A., Sanders, S., Earl, J., O'Neill, S. L., Thomson, N., Sinkins, S. P., & Parkhill, J. (2008). Genome Evolution of Wolbachia Strain wPip from the Culex pipiens Group. *Molecular Biology and Evolution*, *25*(9), 1877-1887.

Characterization of the genomic diversity of *S.* Typhimurium and its monophasic variant in France in pig herds

Madeleine DE SOUSA VIOLANTE[1,2], Carole FEURER[3], Valérie MICHEL[1], Nicolas RADOMSKI[4], Michel-Yves MISTOU[2] and Ludovic MALLET[5]

[1] Actalia, 419 route des champs laitiers, 74800 La Roche sur Foron, France
[2] INRAE, MaIAGE, Université Paris-Saclay, F-78352, Jouy-en-Josas, France
[3] IFIP-Institut du Porc, La Motte au Vicomte B.P. 35104, 35651, Le Rheu Cedex, France
[4] IZSAM, GENPAT, Campo Boario, 64100, Teramo, Italy
[5] Institut Universitaire du Cancer Toulouse - Oncopole, 31059, Toulouse, France

Corresponding Author: madeleine.desousaviolante@anses.fr

Salmonella is one of the most common bacterial pathogen worldwide in human and animal infections, leading to 52,702 cases of human gastroenteritis in Europe in 2020 [1]. To provide new insights on *Salmonella* epidemic investigations, whole genome sequencing (WGS) methods have been developed, especially focusing on the detection of outbreaks and estimation of genetic relationships between isolates [2,3]. *Salmonella* Typhimurium and its monophasic variant are one of the most prevalent serovar [4], and represents 60% of *Salmonella* serovar detected in pig and pork in France [5]. Even if this serovar is well studied, its spreading has not been investigated on the scale of French pig farms, in order to understand whether adaptation mechanisms are linked to geography.

Here, we characterized the genomic diversity of 188 *S.* Typhimurium and its monophasic variant isolated in France from pig herds at the slaughterhouse, using WGS methods. In order to investigate the genome, we detected SNPs using a pangenome-based workflow aiming at evaluating the entire genome of bacterial samples, including accessory genome fractions not shared by all samples. We also performed two approaches aiming at identifying conserved genes and mutations. While the first approach targets known genes based on Abricate software [6] combined with MEGARESv2, Resfinder and VFDB databases, the second approach refers to an in-house algorithm computing sensitivity, specificity and accuracy of accessory genes and core variants according to predefined groups of genomes.

This study brought to light news insights of *S.* Typhimurium and its monophasic variant, which have never been studied at this geographic scale in France. The core and pan-genome phylogenomic analyses revealed a low diversity within monophasic variant of Typhimurium strains, between regions, suggesting a unique clone spreading within pig herds in France. Resistance determinants were found while screening at the gene level, including antibiotics, heavy metals and biocides that could explain the prevalence of these strains within herds. Screening of accessory genes and core variants through an in-house algorithm allowed the identification of conserved mutations to identify genetic markers supporting food safety surveillance without any a-priori. A comparison with monophasic variant isolates from other countries highlighted the genomic specificity of monophasic variants in France, with some exceptions of isolates from bordering countries. This work provides news insights on the dynamics of *S.* Typhimurium and its monophasic variant sampled in pig herds in France.

References

[1] EFSA. The European Union One Health 2020 Zoonoses Report. EFSA Journal, 2021; 19(12):6971
[2] Philip M. Ashton and al. Identification of Salmonella for public health surveillance using wholegenome sequencing. PeerJ4, e1752, 2016
[3] Tim Dallman and al. Phylogenetic structure of European Salmonella Enteritidis outbreak correlates with national and international egg distribution network. Microbial Genomics, 2016
[4] CNR des Escherichia coli, Shigella et Salmonella Pasteur and CHU Robert Debré – APHP. Rapport d'activité annuel 2020 Année d'exercice 2019, 2020.
[5] Leclerc and al. The Salmonella network: a surveillance scheme for Salmonella in the food chain: 2015 results, 2017.
[6] Seeman T. Abricate https://github.com/tseemann/abricate.

Deciphering the distribution of quinone biosynthetic pathways across Proteobacteria

Sophie-Carole CHOBERT[1], Safa BERRAIES[1], Ludovic PELOSI[1], Ivan JUNIER[1], Fabien PIERREL[1], Sophie ABBY[1]

[1] Laboratoire TIMC, UMR 5525, CNRS, Université Grenoble-Alpes, 38000, Grenoble, France

Corresponding Author: sophie-carole.chobert@univ-grenoble-alpes.fr

Isoprenoid quinones are molecules that have a major role in bioenergetics as they shuttle electrons across the respiratory chains of most living organisms. Different types of quinones can be discriminated by their mid-point redox potential. This potential determines respiratory enzymes with which quinones function, accessible respiratory substrates and, ultimately, the environment in which organisms can live.

In Proteobacteria, the two main quinones are menaquinone (MK), a low potential quinone (~-70 mV) and ubiquinone (UQ), a high potential one (~+100 mV). So far, UQ has been considered to be well adapted for aerobic respiration, while MK would rather be involved in anaerobic respiration in O_2-limited contexts. However, our team recently discovered an O_2-independent biosynthetic pathway for UQ production, while the classical pathway depends on the presence of O_2 [1]. It was shown that this O_2-independent pathway is crucial for anaerobic respiration (denitrification) in *Pseudomonas aeruginosa* [2]. Thus, these discoveries challenge the respective assumed physiological roles and origins of MK and UQ pathways in Proteobacteria.

In this study, we systematically investigated quinone production potential across the Proteobacteria phylum. In this prospect, an annotation pipeline assisted by phylogeny was designed in order to infer the presence of the different quinones biosynthetic pathways existing in this phylum (MK, UQ and rhodoquinone (RQ)) and was applied on a large set of over 2500 complete genomes. Particular attention was paid to the genes specific of the UQ O_2-independent pathway, *ubiT*, *-U* and *–V,* which tend to co-localize along genomes. We addressed more particularly the question of their genetic architecture and regulation.

Altogether, this large-scale study gives us more insights to propose an evolutionary scenario of the quinones pathways. It also invites us to revisit the classical view of the respective physiological roles of the respiratory quinones found in Proteobacteria.

References

[1] L. Pelosi *et al.*, « Ubiquinone Biosynthesis over the Entire O2 Range: Characterization of a Conserved O2-Independent Pathway », *mBio*, vol. 10, nº 4, p. 21, 2019.

[2] C.-D.-T. Vo *et al.*, « The O2-independent pathway of ubiquinone biosynthesis is essential for denitrification in Pseudomonas aeruginosa », *Journal of Biological Chemistry*, vol. 295, nº 27, p. 9021-9032, juill. 2020.

Bacterial J-Domain Proteins and Partners Identification and Classification

Roland BARRIOT, Justine LATOUR, Marie-Pierre CASTANIÉ-CORNET, Pierre GENEVAUX and Gwennaele FICHANT

Laboratoire de Microbiologie et Génétique Moléculaires, Centre de Biologie Intégrative (CBI), Université de Toulouse, UPS, France

Corresponding Author: roland.barriot@univ-tlse3.fr

Molecular chaperones maintain cellular protein homeostasis by acting at almost every step in protein biogenesis pathways. In bacteria, the DnaK/HSP70 chaperone has been associated with almost every known essential chaperone functions: it assists the folding of newly synthesized polypeptides, remodels native protein complexes to control their activities, facilitates protein targeting to membranes and protein translocation, reactivates aggregated proteins and participates to protein disaggregation and degradation in cooperation with major disaggregases and proteases. In bacteria, these multifunctional HSP70 chaperones are named DnaK and to act as a *bona fide* chaperone, they strictly rely on essential co-chaperone partners known as the J-domain proteins (JDPs, DnaJ, Hsp40). Genome sequencing has revealed the presence of multiple JDP/DnaK chaperone/co-chaperone pairs in a number of bacterial genomes, suggesting that certain pairs have evolved toward more specific functions. In addition to JDP co-chaperones, most DnaK/HSP70 chaperones also require a nucleotide exchange factor (NEF), named GrpE in *Escherichia coli*. The JDP co-chaperone family is defined by the presence of a compact domain of approximately 70 residues, named the J-domain, which is essential for a functional interaction with DnaK. JDPs can be grouped into three major classes (A, B and C) based on the conserved domains that are present in addition to the J-domain. Adjacent to their N-terminal J-domain, class A JDP members possess a glycine/phenylalanine (G/F)-rich region that connects the J-domain to a zinc-binding domain (ZBD) and a large C-terminal domain that is followed by a short dimerization domain. Class B JDP members have a similar domain architecture, except that they do not have the ZBD. In contrast, class C JDPs only have the J-domain in common with class A and B and which can be found at the N- or at the C-terminal ends, or even in the middle of the JDP architecture.

In this study[1], we implemented a strategy for the identification, classification and phylogenetic analysis of the different partners (JDPs, HSP70 and NEF) among complete bacterial proteomes. A representative set of 1,709 bacterial genomes was selected to maximize the coverage of the bacterial diversity and taxonomy. Fast and sensitive screening of full proteomes to retain the different partners candidates was performed by using HMM of known domains. All InterPro domains of the candidates were retrieved to achieve a finer classification into known classes of HSP70, JDP and NEF in *E. coli*. For HSP70 and JDPs, we designed decision trees based on the combinations of InterPro domains and/or locally built HMMs. HSP70 proteins were classified into DnaK, HscA and HscC. The JDPs were subclassified into class A, class B, class C-DjlA, class C-DjlB/C, class C-HscB, class C-wDomain (harboring some other domains), and class C-woDomain (no other known domain is detected). This confirmed the overall presence of DnaK-DnaJ-GrpE in bacterial genomes, and the well preserved co-presence of HscA with HscB and HscC with DjlB/C. The class C-wDomain proteins were grouped into at least 66 different associations likely representing new groups of class C JDPs spread in different classes of bacteria. For the class C-woDomain proteins, we searched for conserved regions not yet integrated in domain databases. To this end, we performed all against all pairwise sequence comparisons leading to a weighted graph of similar regions. Community detection produced clusters of similar regions. The multiple sequence alignments of these regions were manually inspected and selected for HMM construction. Eventually, we kept 8 conserved regions and proposed them as domains specific to new subclasses of JDPs that need further attention and investigation.

Acknowledgments

We thank Petra Langendijk-Genevaux and Yves Quentin for their valuable advice. Clément Birbes, Callum Burnard and Cyril Kurylo are also acknowledged for their contribution at the beginning of this project as first year master students from the University Paul Sabatier Toulouse. This work was supported by ANR-2019/ChapCop to P.G.

References

1. Roland Barriot, Justine Latour, Marie-Pierre Castanié-Cornet, Gwennaele Fichant and Pierre Genevaux. *J-Domain Proteins in Bacteria and their Viruses.* J. Mol. Biol., 432(13):3771-3789, 2020.

Identification of conserved Regulatory Sequences in Ray-finned Fishes

Jeanne BAUDUIN[1], François GIUDICELLI[1] and Hugues ROEST CROLLIUS[1]

[1] Institut de Biologie de l'ENS (IBENS), Département de biologie, École normale supérieure, CNRS, INSERM, Université PSL ; 46 rue d'Ulm, 75005 Paris, France

Corresponding Author: bauduin@bio.ens.psl.eu

1 Introduction

Non-coding regulatory sequences, such as enhancers or promoters, are essential in all living organisms as they monitor the expression of the genome. Evolution therefore exerts selective pressure on them, leading to conservation of these genomic structures across various species. While this conservation is widely recognized, genome-wide studies of homology relationships between regulatory elements across large panels of species are largely absent in the literature, leading to a gap of knowledge regarding the evolution of enhancers. The Actinopterygii clade (ray-finned fishes) is particularly interesting in this respect, due to several whole genome duplication events (i.e., Teleostea, Salmonidae, Cyprinidae).

We use a combination of conserved sequence and synteny criteria to identify putative regulatory elements and their gene targets among Actinopterygii, as well as the homology relationships between them. This allows us to investigate the evolution of regulatory relationships at an unprecedented level.

2 Results

We identify Conserved Non-coding Elements (CNEs) at the genome scale by integrating the conservation of sequence as detected in multiple alignments with the conservation of synteny between CNEs and neighboring genes.

In the continuation of the work in [1] and [2] (PEGASUS method), we are focusing on the identification of CNEs in 46 ray-finned fish species chosen on the basis of their phylogenetic position and the quality of their genome assembly and annotation. Starting from a homemade multiZ alignment of the 46 genomes, we use a per-base conservation score based on the phylogeny [3] to identify conserved windows outside of annotated exons and repeats. We thus obtain CNEs shared across several species and look for genes that are consistently present in their neighborhood in the different species to identify the most likely target(s) associated to the regulatory sequence. The knowledge of the species in which a CNE is found also allows to retrace its age and evolutionary history.

Our method scans a whole multiple alignment file for CNEs in less than a day. These CNEs show significant overlap with zebrafish ATAC-seq peaks and predicted enhancers in [4], which is consistent with their putative regulatory role. We can attribute an age to each of these CNEs and therefore detect the emergence of particular regulations associated to specific functions along the evolution of fishes.

References

[1] Clément Y, Torbey P, Gilardi-Hebenstreit P, Roest Crollius H. *Enhancer-gene maps in the human and zebrafish genomes using evolutionary linkage conservation*. Nucleic Acids Res. 2020 Mar 18;48(5):2357-2371. doi: 10.1093/nar/gkz1199. PMID: 31943068; PMCID: PMC7049698.

[2] Naville M, Ishibashi M, Ferg M, et al. *Long-range evolutionary constraints reveal cis-regulatory interactions on the human X chromosome*. Nat Commun. 2015;6:6904. Published 2015 Apr 24. doi:10.1038/ncomms7904

[3] Davydov EV, Goode DL, Sirota M, Cooper GM, Sidow A, Batzoglou S. *Identifying a high fraction of the human genome to be under selective constraint using GERP++*. PLoS Comput Biol. 2010 Dec 2;6(12):e1001025. doi: 10.1371/journal.pcbi.1001025. PMID: 21152010; PMCID: PMC2996323.

[4] Yang, H., Luan, Y., Liu, T. et al. *A map of cis-regulatory elements and 3D genome structures in zebrafish*. Nature 588, 337–343 (2020). https://doi-org.insb.bib.cnrs.fr/10.1038/s41586-020-2962-9

MacSyFinder v2: An improved search engine to model and identify molecular systems in genomes

Bertrand Néron[1], Rémi Denise[2], Charles Coluzzi[2], Marie Touchon[2], Eduardo P.C. Rocha[2], Sophie S. Abby[3]

[1] Institut Pasteur, Université de Paris Cité, Bioinformatics and Biostatistics Hub, 75015 Paris, France

[2] Institut Pasteur, Université de Paris Cité, CNRS UMR3525, Microbial Evolutionary Genomics, 75015 Paris, France

[3] Laboratoire TIMC, UMR 5525, CNRS, Université Grenoble-Alpes, 38000, Grenoble, France

Corresponding Author: bneron@pasteur.fr

Complex cellular functions are most often encoded by a set of genes rather than individual ones. Furthermore, the genes in such “systems” are often encoded nearby in microbial genomes. MacSyFinder uses these properties to model and then accurately annotate cellular functions in microbial genomes at the system-level rather than at the individual-gene level. We hereby present a major release of MacSyFinder [1], MacSyFinder version 2 (v2). This new version is coded in Python 3 (>= 3.7). The code was improved and rationalized to enable higher maintainability over time. Several new features were added to allow more flexible modeling of the systems. We introduce a more intuitive and comprehensive search engine to identify all the best candidate systems and sub-optimal ones that still respect the models’ constraints. We also present the novel *macsydata* companion tool that enables the easy installation and broad distribution of the models developed for MacSyFinder (macsy-models) from GitHub repositories. Finally, we have updated, improved, and made available MacSyFinder popular models to this novel version: TXSScan and TFF-SF, CONJscan, and CasFinder. MacSyFinder v2 can be found at this URL: https://github.com/gem-pasteur/macsyfinder

References

1. Sophie S Abby, Bertrand Néron, Hervé Ménager, Marie Touchon, Eduardo PC Rocha. MacSyFinder: A Program to Mine Genomes for Molecular Systems with an Application to CRISPR-Cas Systems. *PLOS ONE*, https://doi.org/10.1371/journal.pone.0110726, 2014.

Structural variation involving transposable elements associated with grain width in cultivated rice.

Marie-Christine Carpentier[1,2], Christel Llauro[1,2], Wan-Yi Qiu[3], Yue-Ie Hsing[3] and Olivier Panaud[1,2]

[1] Centre National de la Recherche Scientifique (CNRS), Laboratory of Plant Genome and Development. UMR 5096, 66860 Perpignan, France

[2] University of Perpignan, Laboratory of Plant Genome and Development. UMR 5096, 66860 Perpignan, France

[3] Institute of Plant andMicrobial Biology, Academia Sinica, Nankang 115 Taipei, Taiwan

Corresponding author: marie-christine.carpentier@univ-perp.fr

Rice, *Oryza sativa*, is the staple food for half the world population. It is the first crop whose genome has been sequenced in 2005. This model species benefits from various genomic ressources among which genome sequences of 3000 rice varieties that were publicly released in 2014 [1]. This provides a unique opportunity to unravel the genetic diversity of the crop with accessions from 89 countries, distributed into 5 varietal groups – *indica, japonica, aus/boro, basmati/sandri* and *intermediate*.

Transposable elements (TE) are mobile genetic elements abundant in plant genomes. Knowledge of their impact on the structure, function and evolution of genome, these mobile entities can provide a more precise picture of rice genome dynamics on a shorter evolutionary scale (posterior to the domestication) because their transposition rate is higher than base substitutions.

We have developped a pipeline, nammed TRACKPOSON, to detect all retrotransposons insertions in the 3000 genomes dataset [2]. With these results, to understand the funtional impact of TE on this crop, we performed a genome-wide association study (TE-GWAS, [3]) with different agronomic traits. We found a significant association between an insertion of TE and rice grain width. If the TE is present, the grain is larger.

For further analysis, we sequenced 2 phylogenetically related rice varieties (one thin grain and one large grain) with Nanopore technologies. Thanks to long-read sequencing, after assembly and genomic analysis, we validated the insertions of TE. In addition, we observed that the insertion region is part of a larger insertion, possibly an introgression from another rice variety (appears to be an ancestral wild rice). Analyses of introgression are underway to annotate the genes, insertions of TE insertion and genomic comparision between the 2 cultivated rice varieties.

In parallel, genetic analysis is underway by crossing the two rice plants together and creating a rice population for futures analysis.

References

[1] The 3,000 rice genomes project. The 3,000 rice genomes project. *GigaScience*, 3(1):7, December 2014.

[2] Marie-Christine Carpentier, Ernandes Manfroi, Fu-Jin Wei, Hshin-Ping Wu, Eric Lasserre, Christel Llauro, Emilie Debladis, Roland Akakpo, Yue-Ie Hsing, and Olivier Panaud. Retrotranspositional landscape of Asian rice revealed by 3000 genomes. *Nature Communications*, 10(1):24, December 2019.

[3] Roland Akakpo, Marie-Christine Carpentier, Yue Ie Hsing, and Olivier Panaud. The impact of transposable elements on the structure, evolution and function of the rice genome. *New Phytologist*, 226(1):44–49, April 2020.

Evolution of Tandemly Arrayed Genes in Rosaceae

Martin LEDUC [1], Tanguy LALLEMAND [1], Carène RIZZON[2], Jérémy CLOTAULT[1], Jean-Marc CELTON[1] and Claudine LANDES[1]

[1] Univ Angers, Institut Agro, INRAE, IRHS, SFR QUASAV, F-49000 Angers, France
[2] Laboratoire de Mathématiques et Modélisation d'Evry (LaMME), Université d'Evry Val d'Essonne, Université Paris-Saclay, UMR CNRS 8071, ENSIIE, USC INRAE, 23 bvd de France, CEDEX, 91037 Evry Paris, France
Corresponding Author: claudine.landes@inrae.fr

Since Susumu Ohno's hypothesis in 1970 [1], it is commonly accepted that gene duplications have an essential role in the evolution of organisms and that this mechanism is a key for genetic innovations. There are several mechanisms within genomes that generate large-scale duplications (Whole Genome Duplication, Segmental Duplication) or small-scale duplications (Tandem Duplication of genes). After duplication the main fate of duplicated genes is loss by pseudogenization and/or chromosomal remodeling, however there are other evolutionary scenarios that retain duplicated genes: dosage effect, cellular or tissue adaptation, sub-functionalization and neo-functionalization. However, the mechanisms that control the fate of duplicated genes are still largely unknown. We are interested in this question of the evolution of duplicated genes in rosaceous plants, more specifically in apple and rose for which we have participated in the sequencing of a high quality genome [2-3]. The question we address here is: Is the selection pressure on tandemly duplicated genes the same regardless of the size of the gene cluster and is it the same for recent or old duplications?

Using OrthoFinder [4], we determined all gene families within seven Rosaceae species for which a quality genome is available: Strawberry (*Fragaria vesca* – diploid-*, Fragaria x ananassa* – octoploid -), Bramble (*Rubus occidentalis*), Rose (*Rosa chinensis*), Apple (*Malus x domestica*), Pear (*Pyrus communis*), Peach (*Prunus persica*) with Arabidopsis (*Arabidopsis thaliana*) as an outgroup. Then, using the i-ADHoRe software [5], we assigned to each gene of these multigene families their duplication status: TAG (Tandemly Arrayed Genes), SD (Segmental Duplication), dispersed (other duplication types).

This unbiased analysis highlighted species specificities concerning the nature of gene duplication: for example, apple, pear and *F. ananassa* have a very high number of SDs which is expected since the first two have undergone a recent Whole Genome Duplication (50 Mya) and *F. ananassa* is an octoploid (the only polyploid in our data set). In contrast, the rose has a very large number of TAGs. As the rose is a species of interest to our institute (as well as apple and pear) we decided to further explore this particularity of the rose.

We then used PAML [6] to calculate the Ka/Ks evolutionary rates using the model YN00 [7] for all duplicated gene pairs and analyzed their distribution according to the size of the TAG clusters and the estimated age of the duplications. However, the mechanism of gene conversion introduces a bias in the estimation of Ks. To get rid of this confounding factor we verified our results (only in apple) for which we gave the status of young TAG to the clusters that exist only in one of the post-WGD syntenic fragment, those present on both are being considered old.

References

1. Ohno S. *Evolution by gene duplication.* Springer-Verlag, 1970.
2. Daccord N, Celton J-M *et al*. *High quality de novo assembly of the Apple genome and methylome dynamics of early fruit development*. Nature Genetics, 2017, 49: 1099-1106.
3. Hibrand Saint-Oyant L, *et al. A high-quality genome sequence of Rosa chinensis to elucidate ornamental traits*. Nature Plants, 2018, 4: 473-484.
4. Emms DM & Kelly S. *OrthoFinder:Solving fundamental biases in whole genome comparisons dramatically improves orthogroup inference accuracy.* Genome Biology, 2015, 16, 157.
5. Proost S *et al. i-ADHoRe3.0—Fast and sensitive detection of genomic homology in extremely large data sets*. Nucleic Acids Res. 2012, 40, e11.
6. Yang, Z. *PAML 4: Phylogenetic analysis by maximum likelihood*. Mol. Biol. Evol., 2007, 24, 1586–1591.
7. Yang & Nielsen, *Estimating synonymous and nonsynonymous substitution rates under realistic evolutionary models.* Mol. Biol. Evol., 2000, 17, 32-43

Telomere-to-telomere genome assembly of the phytoparasitic nematode and virus vector *Xiphinema index*

Karine Robbe-Sermesant[1], Arthur Péré[1], Corinne Rancurel[1], Christophe Klopp[2], Laetitia Perfus-Barbeoch[1], Daniel Esmenjaud[1], Cyril Van Ghelder[1], Marie Gislard[3], Céline Lopez-Roques[3], Carole Iampietro[3], Ulysse Portier[1], Etienne GJ Danchin[1]

1 ISA, Université Côte d'Azur, INRAE, CNRS, 400 route des chappes, 06903, Sophia Antipolis, France

2 SIGENAE, MIAT, INRAE, 24 chemin de Borde-Rouge Auzeville, 31326, Castanet-Tolosan, France

3 GeT-PlaGe, 24 chemin de borde rouge Auzeville, 31326, Castanet-Tolosan, France

Corresponding Author: etienne.danchin@inrae.fr

Xiphinema index, in the order of *Dorylaimida*, is a Plant Parasitic Nematode (PPN) that feeds through root tips. *X index* is the vector of Grapevine FanLeaf Virus (GFLV) causing major damages in vineyards worldwide. Reproduction is mostly by meiotic parthenogenesis (male and outcrossing events are very rare). A single juvenile can yield a whole population. The *X. index* genome is described as diploid with two pairs of 10 chromosomes. Flow cytometry experiments have estimated a haploid genome size of ca. 205Mb.

High quality and long PacBio HiFi reads (N50: 13.6 kb, longest: 40.3 kb, volume: 8.4 Gb, coverage: 36X) were used for a primary assembly using HiFiasm. This yielded 172 contigs with a total assembly size of 260Mb and a N50 of 12.6Mb. This initial assembly was bigger than the estimated genome size by both flow cytometry and k-mer. Using contact maps from Hi-C data (Arima protocol, Illumina paired reads), we obtained a final 221 Mb haploid genome. Contact maps were produced using Juicer. Final manual scaffolding was realized using JuiceBox with the help of nematode specific telomere motifs identified using TIDK. Assembly quality assessment was performed using KAT, BUSCO, and Blobtools. A 14.8 kb mitochondrial genome was assembled using an in-house tool mitochondriAL circulAr DNA ReconstItutioN (ALADIN). Several transfers of mitochondrial genetic material into the nuclear genome were detected and supported by overlapping long reads, the largest insertion being 11.4 kb long.

It is the first telomere to telomere assembled nematode genome in the *Dorylaimida* order and a breakthrough for research on vineyards protection. This highly contiguous and complete genome opens many perspectives for comparative genomics and discovery of the key elements involved in the evolution of plant parasitism.

Gene orthology detection for Long Non Coding RNA (LncRNA)

Fabien Degalez[1] and Sandrine Lagarrigue[1]

[1] INRAE, INSTITUT AGRO, PEGASE UMR 1348, 35590, Saint-Gilles, France

Corresponding Authors: fabien.degalez@inrae.fr sandrine.lagarrigue@agrocampus-ouest.fr

1. Introduction

A major component of the transcribed genomes, discovered in the last decade is long non-coding RNAs (lncRNAs), which are defined as transcripts of more than 200 nucleotides with low potential coding capabilities. Such lncRNAs increase the list of regulatory elements of the gene expression. However, the role of most of them remains to be clarified. Exploring the lncRNA conservation between species is an approach to strengthen the annotation of lncRNAs by *i)* reinforcing the existence of these lncRNAs (most of them are gene prediction and are weakly expressed) and *ii)* making possible function inference in a species from another one more studied such as human or mouse as previously done for protein coding genes (PCGs). However, contrary to the PCGs, the primary sequences of lncRNAs are not well conserved between species; LncRNAs are conserved in patches of a few nucleotides (5 to 30) with an arrangement of these patches that is not necessarily conserved [1]. Therefore, the usual reference databases such as Ensembl BioMart, do not provide any information about lncRNA orthology whatever the species while e.g. for chicken 72% (12,050 among 16,779 e!106) of the PCGs are “1 to 1” orthologous with human. A few specialized databases reported lncRNAs orthologs such as SyntDB [2] which focuses on the ape group or NONCODEv6 [3] which targets 12 species from different branches. Both used whole-genome alignment approaches based on syntenic regions identified with liftOver, and with reduced gap penalties. In this context, we have developed a workflow combining different approaches which can be used for any species of interest and applied it on 8 species covering a broad phylogenetic scale from mammals to chicken.

2. Results

The workflow combines three methods: two methods based on synteny established for PCGs and a third based on block alignment. The method 1 uses a “PCG-lncRNA-PCG” triplet with orthologous PCGs on either side of the lncRNA as an "anchor". The method 2 is based on “lncRNA-PCG” pairs in which the PCG has to be orthologous in both species and the lncRNA PCG transcripts have to be in the same genomic orientation in both species. Finally, the last method considers the alignment of lncRNAs but by small patches, using the “Mercator-Pecan” multiple genome alignment method.

Applied on human, mouse, pig, cow, goat, dog, horse and chicken, the three methods are complementary regardless of the species pair. Considering the 18,805 human lncRNAs (e!106), around 10,000 are identified with one orthologous neighboring PCG in chicken among them more than 1,000 lncRNA loci can be considered as high quality orthologs because they are detected by the three methods. Moreover, more than 2,000 lncRNAs are detected by the first and second method. To test the relevance of these lncRNAs, different analyses based on the conservation of the co-expression network between species have been performed. The first results show interesting cases of network conservation between human and chicken, whose functional term enrichment is in progress.

Aknowledgments

We thank Thomas Derrien (CNRS, IGDR, Rennes), Sylvain Foissac (INRAE, UMR GenPhySe, Toulouse) and Hervé Acloque (INRAE – UMR GABI, Paris) for discussions and advice.

References

1. T Noviello et al. BMC Bioinformatics 19, 407 (2018). https://doi.org/10.1186/s12859-018-2441-6
2. Bryzghalov et al. Nucleic Acids Research 48 D238-45 (2020). https://doi.org/10.1093/nar/gkz941.
3. Zhao et al. Nucleic Acids Research 49 (D1): D165-71 (2021). https://doi.org/10.1093/nar/gkaa1046.

Exploration and modeling the evolution of metabolic networks in fungi

Vahiniaina ANDRIAMANGA[1] , Anne LOPES[1] and Olivier LESPINET[1]

[1] Institute for Integrative Biology of the Cell (I2BC), 1 avenue de la terrasse bâtiment 21 , 91190 , Gif-Sur-yvette , France

Corresponding Author: vahiniaina.andriamanga@i2bc.paris-saclay.fr

Background: Metabolism is a set of biochemical reactions that take place inside an organism. The metabolic network represents the relationships between all these biochemical reactions and defines the metabolic capacity of the organism; particularly, the capacity to use compounds in the medium and/or to synthesize new products. The evolution of metabolic networks is subjected to the constraints exerted by the environment. Consequently each organism has a specific metabolic profile that illustrates its ability to transform chemical compounds in the relationship of the species with its environment. In order to better understand how the environment has shaped the evolution of the metabolic network(s), we investigated the evolution of the nodes (enzymes) but also the edges (shared compounds between two enzymes) of the metabolic networks of 175 fungal species by combining graph based analyses and comparison of enzyme conservation profiles. Fungi are known to have a wide variety of metabolic profiles thereby constitute a very good model to address this question. This integrative approach combined with phylogenetic information provides an opportunity to decipher the evolution of the metabolism from the network level to enzyme level.

Results: With phylostratigraphy approaches, we dated the origin of the metabolic enzymes of 175 fungal species. We showed that 708 out of 954 enzymes were already present at the origin of fungi. Interestingly, we show that only half of them has been conserved in all the considered species while the other half is only present in specific lineages or subtrees suggesting multiple lost events in their evolutionary history. We then investigated the conservation level of enzymes with respect to their localization with topological metrics and showed that the enzymes conserved in fungi display a higher centrality (i.e., tend to be located at the center of the network) than those that are subtree or lineage-specific. In addition, conserved enzymes display a higher degree, sharing more compounds with other enzymes than non-conserved ones. The metabolic network can be divided into pathways (series of enzymes that lead to usable materials). Our analysis enables the distinction between pathways common to all fungal species and those that are specific to subsets of species. Specific pathways (i) are mainly involved in accessory metabolism (ii) are generally located at the periphery of the network and (iii) have less connections with other pathways.Contrarily, pathways common to all fungal species (i) are characterized by long sequences of reactions, involving subsets of highly conserved reactions, (ii) are generally located at the center of the network and (iii) are highly connected to other pathways.

Conclusions: Our analysis shows that (i) half of the metabolic enzymes are highly conserved and are likely to be ancient and (ii) the evolution of the metabolic network was mostly driven by many loss of enzymes in specific lineage or subtrees. Non-conserved enzymes (idem, specific pathways) are characterized by specific topological metrics (i.e., centrality measure and degree) that may explain the fact that they could have been lost or gained in specific lineage.

DNA methylation patterns of transcription factor binding regions characterize their functional and evolutionary contexts

Martina RIMOLDI[1], Duncan ODOM[2,3], Jussi TAIPALE[4,5,6], Paul FLICEK[1,2,7] and Maša ROLLER[1]

[1] European Molecular Biology Laboratory, European Bioinformatics Institute, Wellcome Genome Campus, Hinxton, Cambridge, CB10 1SD, UK

[2] Cancer Research UK Cambridge Institute, University of Cambridge, Robinson Way, Cambridge, CB2 0RE, UK

[3] German Cancer Research Center (DKFZ), Division of Regulatory Genomics and Cancer Evolution, Im Neuenheimer Feld 280, Heidelberg 69120, Germany

[4] Division of Functional Genomics and Systems Biology, Department of Medical Biochemistry and Biophysics, Karolinska Institutet, SE 141 83 Stockholm, Sweden

[5] Genome-Scale Biology Program, Post Office Box 63, FI-00014 University of Helsinki, Helsinki, Finland

[6] University of Cambridge, Department of Biochemistry, Tennis Court Road, Cambridge, CB2 1QW, UK

[7] Wellcome Sanger Institute, Wellcome Genome Campus, Hinxton, Cambridge, CB10 1SD, UK

Corresponding author: rimoldi@ebi.ac.uk

DNA methylation is an important epigenetic modification which has numerous roles in modulating genome function. Its levels are spatially correlated across the genome, typically high in repressed regions but low in transcription factors (TF) binding sites and active regulatory regions. However, the mechanisms establishing genome-wide and TF binding site methylation patterns are still unclear.

We used a comparative approach to investigate the association of DNA methylation to TF binding evolution in mammals. Specifically, we experimentally profiled DNA methylation and TF occupancy of five distinct TFs (CTCF, CEBPA, HNF4A, HNF6, FOXA1) in the liver of five mammalian species (human, macaque, mouse, rat, dog). TF binding sites were lowly methylated, but they often also had intermediate methylation levels. Employing a classification and clustering approach, we extracted distinct patterns of DNA methylation levels at TF bound regions that were conserved across species. CEBPA, HNF4A, HNF6 and FOXA1 shared the same methylation patterns, while CTCF's differ. These patterns characterize distinct functions and chromatin landscapes of regions where the TFs bind. Leveraging our phylogenetic framework, we found DNA methylation gain upon evolutionary loss of TF occupancy, indicating their coordinated evolution. Furthermore, DNA methylation turnover differs between distinct methylation patterns, reflects their genomic contexts, and may be subject to distinct evolutionary forces.

Our comparative epigenomic analyses reveal important insights in DNA methylation features of TF binding, their associations to regulatory activity, chromatin contexts, and epigenome evolution in mammals.

Multispecies comparison of fruit development through mRNA quantification analysis

Chloé Beaumont[1], Sylvain Prigent[1,2], Yves Gibon[1,2] and Sophie Colombié[1,2]

[1] INRAE, Univ. Bordeaux, UMR1332 BFP, 33882 Villenave d'Ornon, France
[2] Bordeaux Metabolome, MetaboHUB, PHENOME-EMPHASIS, 33140 Villenave d'Ornon, France

Corresponding Author: chloe.beaumont@inrae.fr

Fruit development is a highly complex continuous process divided in several stages to reach a ripe fruit. Gene expression changes during fruit development and ripening to adapt to any special needs of fruits. Transcriptomics is an effective tool to study the functional reprogramming of cells. Fruits have a diversity of types adapted to protect the seed during its development and to allow its dissemination. However, the mechanisms underlying the diversity of fruit types are not well understood.

Combining high throughput next generation sequencing applied on transcriptomic data and the addition of quantitative spikes in the experimental design [1], we were able to determine in a quantitative manner transcript concentrations along nine developmental stages for seven different fruits : tomato, pepper, eggplant, kiwifruit, cucumber, apple and strawberry. We investigated the molecular events that correlate with fruit development using general statistics comparison and clustering analysis. The study of the functional annotation of transcripts predominantly expressed during the different fruits development stages allowed us to determine genes involved in those stages as ripening.

Although transcriptomics allowed us to study gene expression, protein abundance is largely affected by post-transcriptional and post-translational regulations. In the tomato fruit, mRNA and protein levels are poorly correlated throughout development [2]. A mathematical model of protein translation based on an ordinary differential equation involving synthesis and degradation rate constants have been developed to estimate the stability of protein by combining transcriptomic and proteomic data [2]. Adding proteomic data to this transcriptomic dataset on the studied fruits will allow the computation of synthesis and degradation rates for proteins in those seven different fruits, hence the comparison of protein stability between fruits.

References

[1] Belouah, I.; Bénard, C.; Denton, A.; Blein-Nicolas, M.; Balliau, T.; Teyssier, E.; Gallusci, P.; Bouchez, O.; Usadel, B.; Zivy, M.; et al. Transcriptomic and proteomic data in developing tomato fruit. Data Brief 2020, 28, 105015.

[2] Belouah, I.; Nazaret, C.; Pétriacq, P.; Prigent, S.; Bénard, C.; Mengin, V.; Blein-Nicolas, M.; Denton, A.K.; Balliau, T.; Augé, S.; et al. Modeling Protein Destiny in Developing Fruit. Plant Physiol. 2019, 180, 1709–1724.

MockVirus: expanding viral phylogenetic trees by protein sequence simulation

Julia Kende[1], Thomas Bigot[1], Sarah Temmam[2], Philippe Pérot[2], Béatrice Regnault[2] and Marc Eloit[2]

[1] Institut Pasteur, Université Paris Cité, Bioinformatics and Biostatistics Hub, Paris, France
[2] Institut Pasteur, Université Paris Cité, Pathogen Discovery Laboratory, Paris, France

Corresponding author: Thomas.Bigot@pasteur.fr

Abstract

Generating vast amounts of realistic biological sequences is a powerful way to evaluate and validate bioinformatic methods. Moreover, to better anticipate potential future pandemics, there is a need to improve our knowledge of the virosphere diversity with metagenomic tools able to detect unknown viruses based only on data from distant clades. Simulating realistic virus protein sequences can help us developing and testing such tools.

Following the development of advanced methods for inferring substitution matrix [1,2] and complex substitution models [3,4], several protein sequence simulation tools have been developed [5,6]. However, simulated data usually mimic existing ones, populating a given phylogenetic tree starting from one input root sequence or a randomly generated one, without retaining information from the original alignment apart from those contained in the substitution model. As a result, although simulated sequences can be considered as realistic and representative of a clade in terms of evolution parameters, they could not be considered as belonging to the clade.

MockVirus combines existing tools completed with Python scripts for populating an existing tree with new branches. It takes advantage of the new functionalities offered by the last IQ-TREE version [7]. The ModelFinder tool [8] is first used to choose the best substitution matrix describing the whole sequence evolution. Then, we transform a gamma-rate heterogeneity profile into a partition model which in turns allows us to employ the Alisim tool [6] for simulating new sequences with real amino acid position-specific rates. Simulations are performed on 3-leaf trees with a root sequence representing selected node sequences from the real tree. Distances are dynamically increased along branch lengths with a synchronous BLAST analysis. Insertions and deletions are tackled by simulating the longest length observed in the alignment and randomly recreating gap portions from the initial alignment.

MockVirus has been tested on the *Sarbecovirus* spike protein, aiming at reproducing the emergence of new branches with topological connections similar to the one of the SARS-CoV-2 sub-tree.

References

[1] William R. Pearson. Selecting the Right Similarity-Scoring Matrix. *Current Protocols in Bioinformatics*, 43(1), October 2013.

[2] Bui Quang Minh, Cuong Cao Dang, Le Sy Vinh, and Robert Lanfear. QMaker: Fast and Accurate Method to Estimate Empirical Models of Protein Evolution. *Systematic Biology*, 70:15, 2021.

[3] Le Si Quang, Olivier Gascuel, and Nicolas Lartillot. Empirical profile mixture models for phylogenetic reconstruction. *Bioinformatics*, 24(20):2317–2323, October 2008.

[4] Olga Chernomor, Arndt von Haeseler, and Bui Quang Minh. Terrace Aware Data Structure for Phylogenomic Inference from Supermatrices. *Systematic Biology*, 65(6):997–1008, November 2016.

[5] W. Fletcher and Z. Yang. INDELible: A Flexible Simulator of Biological Sequence Evolution. *Molecular Biology and Evolution*, 26(8):1879–1888, August 2009.

[6] Nhan Ly-Trong, Suha Naser-Khdour, Robert Lanfear, and Bui Quang Minh. AliSim: A Fast and Versatile Phylogenetic Sequence Simulator For the Genomic Era. preprint, Bioinformatics, December 2021.

[7] Bui Quang Minh, Heiko A Schmidt, Olga Chernomor, Dominik Schrempf, Michael D Woodhams, Arndt von Haeseler, and Robert Lanfear. IQ-TREE 2: New Models and Efficient Methods for Phylogenetic Inference in the Genomic Era. *Molecular Biology and Evolution*, 37(5):1530–1534, May 2020.

[8] Subha Kalyaanamoorthy, Bui Quang Minh, Thomas K F Wong, Arndt von Haeseler, and Lars S Jermiin. ModelFinder: fast model selection for accurate phylogenetic estimates. *Nature Methods*, 14(6):587–589, June 2017.

Impact of sequencing platforms on cgMLST and wgSNP analyses on species of Listeria and Salmonella

Yao Amouzou, Norman Wiernasz, Younous Adrouji, Sébastien Leuillet

Biofortis, 3 route de la Chatterie, Saint Herblain, 44800, France

Corresponding Author: yao.amouzou@biofortis.fr

Introduction

Core genome Multilocus Sequence Typing (cgMLST) and whole genome Single Nucleotide Polymorphism (wgSNP)[1] typing are powerful approaches for isolated bacteria analysis during tracking and characterization of foodborne pathogens in order to determine strains relatedness. These analyses are performed using Next Generation Sequencing (NGS) data. However, several NGS technologies are available and their choice could have an impact on obtained results. The aim of this work was to benchmark the cgMLST and wgSNP typing results obtained using different sequencing platforms.

Methods

This study evaluated the effect of three different sequencing platforms on the cgMLST and wgSNP analysis for selected bacterial species. DNA samples from *Salmonella enterica* and *Listeria monocytogenes* were sequenced by short reads (Illumina iSeq 100 and MGI DNBSEQ-G400) and long-reads approaches (Oxford Nanopore MinION Mk1C). The obtained raw sequencing reads were analyzed using BioNumerics 7.6[2] (Applied Maths) with dedicated workflows for cgMLST and wgSNP calling. For comparison, hierarchical clustering was performed based on dissimilarities among allelic profiles for cgMLST and based on sequence composition for wgSNP.

Results

As expected, the raw sequencing reads obtained by the three NGS platforms showed differences in several metrics such as number of raw reads, quality score, reads coverage, reads length, ….

However, similar results were observed on wgSNP (<3 SNPs between sequencing replicates) and cgMLST (same sequence type, >99% of similarity percentage) for each sample across the different NGS platforms, for the two species of interest.

Conclusion

The aim of the study was to evaluate the effect of different sequencing platforms on the cgMLST and wgSNP analysis results. This study suggests that the chosen sequencing platform does not have a significant impact on the analysis to determine strain relatedness for *Listeria monocytogenes* and *Salmonella enterica* species. Thus, the choice of the sequencing platform can rely on external factors such as sequencing cost, multiplexing capacities and turnaround time.

References

1. Jagadeesan B, Gerner-Smidt P, Allard MW, Leuillet S, Winkler A, Xiao Y, Chaffron S, Van Der Vossen J, Tang S, Katase M, McClure P, Kimura B, Ching Chai L, Chapman J, Grant K. The use of next generation sequencing for improving food safety: Translation into practice. Food Microbiol. 2019 Jun;79:96-115. doi: 10.1016/j.fm.2018.11.005. Epub 2018 Nov 17.

2. https://www.applied-maths.com/bionumerics

Automatization and optimization of TEFLoN, an accurate tool for detecting insertions of transposable elements

Corentin Marco[1,2*], Michael C. Fontaine[2,3,4] and Anna-Sophie Fiston-Lavier[1,5]
[1] ISEM, Univ Montpellier, CNRS, UM, Montpellier, France
[2] MIVEGEC, Univ Montpellier, CNRS, IRD, Montpellier, France
[3] Groningen Institute for Evolutionary Life Sciences (GELIFES), Univ Groningen, Groningen, Netherlands
[4] Montpellier Ecology and Evolution of Disease Network (MEEDiN), Montpellier, France
[5] Institut Universitaire de France (IUF)

Corresponding author: corentin.marco@etu.umontpellier.fr* & anna-sophie.fiston-lavier@umontpellier.fr

Transposable elements (TEs) are mobile, repetitive and mutagenic elements of DNA known to be important component of eukaryote genomes and major actors in genome evolution. TEs can create genetic novelty, diversity, and have been involved in evolutionary processes such as adaptation [1]. In order to estimate their impact on genome evolution and adaptive processes, it is necessary to study their dynamics. We need first to detect them from individual to populations. Due to their repetitive nature, detection of TE insertions using paired-end reads revealed a high false-positive rate, up to 40%. One solution suggested is to combine the results from up to three distinct tools relying on different approaches [2]. One of the most promising tools highlighted in such study is TEFLoN that provides the most accurate results for all TE families, regardless of the data quality [3]. This tool developed in python2.7 is easy to install and use, thanks to its low number of dependencies. TEFLoN takes as input short paired-end reads, a reference genome, and its TE annotations and/or a library of TE consensi. It is composed of four main steps to (1) detect all TE insertions, (2) filter out the TE insertions with low quality evidences at the individual and population levels, (3) genotype each selected TE insertions and (4) estimate the allele frequency for each TE insertions.
However, many technical limitations have been identified : Each script must be launched independently without parallelisation that makes it time and memory consuming. Moreover, a large number of files are created. We propose here to develop TEFLoN2 to make the TE detection and analysis in sequencing data faster and handling large datasets, such as those of consortium resequencing efforts [4] [5]. To do so, the codes had been updated to python3.X and a Snakemake pipeline developed [6] to automatize and parallelize the TEFLoN steps [3]. To facilitate the TEFLon distribution, we are now working on a Singularity containerized system [7].
A preliminary benchmark between TEFLoN and TEFLoN2 using population resequencing data from the *Anopheles gambiae* 1000 genome consortium [4] [5] suggests a clear decrease in computing time with a constant accuracy of the TE detection results.

References

[1] Zi Wen Li, Xing Hui Hou, Jia Fu Chen, Yong Chao Xu, Qiong Wu, Josefa González, and Ya Long Guo. Transposable elements contribute to the adaptation of arabidopsis thaliana. *Genome Biology and Evolution*, 10(8):2140–2150, 2018.

[2] Vendrell-Mir P., Barteri F., Merenciano M., González J., Casacuberta J.M., and Castanera R. A benchmark of transposon insertion detection tools using real data. *Mobile DNA*, 10(1):53, 2019.

[3] Adrion J.R., Song M.J., Schrider D.R., M.W. Hahn, and Schaack S. Genome-wide estimates of transposable element insertion and deletion rates in drosophila melanogaster. *Genome Biology and Evolution*, 9(5):1329–1340, 2017.

[4] The Anopheles gambiae 1000 Genomes Consortium. Genetic diversity of the african malaria vector anopheles gambiae. *Nature*, 552(7683):96–100, 2017.

[5] The Anopheles gambiae 1000 Genomes Consortium. Genome variation and population structure among 1142 mosquitoes of the african malaria vector species *Anopheles gambiae* and *Anopheles coluzzii*. *Genome Research*, 30(10):1533–1546, 2020.

[6] F Mölder, KP Jablonski, B Letcher, MB Hall, CH Tomkins-Tinch, V Sochat, J Forster, S Lee, SO Twardziok, A Kanitz, A Wilm, M Holtgrewe, S Rahmann, S Nahnsen, and J Köster. Sustainable data analysis with snakemake. *F1000Research*, 10(33), 2021.

[7] Gregory M. Kurtzer, Vanessa Sochat, and Michael W. Bauer. Singularity: Scientific containers for mobility of compute. *PLOS ONE*, 12(5):1–20, 2017.

Fast Construction & Extension of Gene Families

Aurélie Maurin[1], Franklin Delehelle[1], Alexandra Louis[1] and Hugues Roest Crollius[1]
Institut de Biologie de l'ENS (IBENS), Département de biologie, École normale supérieure, CNRS, INSERM, Université PSL; 46 rue d'Ulm, 75005 Paris, France

Corresponding author: maurin@bio.ens.psl.eu

1 Introduction

An essential step in comparative genomics analyses is the correct creation of homologous gene families. The increasing flow of new genomes being sequenced ever faster and more accurately across the tree of life [1] suggests fascinating prospects for this type of analysis, especially for exploring new functions and modes of evolution. Gene families characterization is a computationally intensive process, subject to many more or less arbitrary choices. Often, to get around the problem of computation time consumption, only a small subset of the available genomes is typically included in the definition of theses families, which implies loss of information. It therefore becomes clear that this influx of data create a challenging situation regarding scaling, requiring bioinformatics tools of adequate performances. In this poster, we present an approach whose objective is, on one side, to target a set of genomes for the initial definition of families an order of magnitude larger than currently, and, on the other side to progressively increment the families as new genomes become available, reducing the need for *de novo* reconstructions.

2 Results

We tested different similarity search and sequence alignment methods [2,3,4] to benchmark and extrapolate their time and memory consumption. We then compared several algorithms for recreating gene families, based both on sequence and synteny, either *ex nihilo* [5,6] or incremental [7]. Finally, we combined these recent and powerful algorithms to create a hybrid and reproducible pipeline following the FAIR principles and aiming at both (i) creating gene families *ex nihilo*, and (ii) extending them incrementally – and thus rapidly – without sacrificing accuracy. We ensured the credibility of our results by keeping the methods that maximize the number of families whose size matches the number of genomes chosen. We present our results in terms of algorithm choice, computation time and memory cost for the definition of gene families on catalogs of genomes of increasing sizes, allowing us to extrapolate the performance in terms of scaling up.

References

[1] Mark Blaxter, John M Archibald, Anna K Childers, Jonathan A Coddington, Keith A Crandall, Federica Di Palma, Richard Durbin, Scott V Edwards, Jennifer AM Graves, Kevin J Hackett, et al. Why sequence all eukaryotes? *Proceedings of the National Academy of Sciences*, 119(4), 2022.

[2] Benjamin Buchfink, Chao Xie, and Daniel H Huson. Fast and sensitive protein alignment using diamond. *Nature methods*, 12(1):59–60, 2015.

[3] W James Kent. Blat—the blast-like alignment tool. *Genome research*, 12(4):656–664, 2002.

[4] Martin Steinegger and Johannes Söding. Mmseqs2 enables sensitive protein sequence searching for the analysis of massive data sets. *Nature biotechnology*, 35(11):1026–1028, 2017.

[5] Leland McInnes, John Healy, and Steve Astels. hdbscan: Hierarchical density based clustering. *J. Open Source Softw.*, 2(11):205, 2017.

[6] JM Kinet and MM Peet. Tomato. Technical report, CABI, 1997.

[7] Victor Rossier, Alex Warwick Vesztrocy, Marc Robinson-Rechavi, and Christophe Dessimoz. Omamer: tree-driven and alignment-free protein assignment to subfamilies outperforms closest sequence approaches. *Bioinformatics*, 37(18):2866–2873, 2021.

Caulifinder : a pipeline for automatic detection and annotation of endogenous viral sequences of *Caulimoviridae*

Helena Vassilieff[1], Véronique Jamilloux[1], Sana Haddad[1], Nathalie Choisne[1], Vikas Sharma[1*], Pierre-Yves Teycheney [2] and Florian Maumus [1]

[1] Université Paris-Saclay, INRAE, URGI, 78026, Versailles, France.

[2] CIRAD, UMR PVBMT, F-97410 Saint Pierre, La Réunion, France

[*]Present address : Institute of Bio- and Geosciences, IBG-1, Biotechnology, Forschungszentrum Jülich GmbH, Jülich, Germany

Corresponding Author: florian.maumus@inrae.fr

Endogenous viral sequences (EVEs) result from the active (by a viral enzyme) or passive (by non-homologous recombination) integration of full or part of viral genomes into the genome of eukaryotic or prokaryotic hosts. EVEs can represent a significant part of the host's genome. They provide access to often ancient sequences that can be used in paleovirology approaches to study the evolution of viruses over time steps through several million years [1,2]. However, EVEs are frequently ignored in genome annotation due to the lack of dedicated bioinformatics tools for their detection and characterisation, which have long relied on manual processes. In plants, most of the known EVEs belong to the family *Caulimoviridae* [3], the only family of retrotranscribed plant viruses [4].

Here we present "Caulifinder", a bioinformatics pipeline for automatic characterisation and annotation of *Caulimoviridae* EVEs in plant genomes. Caulifinder consists of two complementary pipelines. The first pipeline detects *Caulimoviridae* EVEs and performs an automatic reconstruction of consensus sequences. To perform this, it uses elements of the REPET suite [5,6]. It uses these consensus sequences to automatically annotate the Caulimoviridae EVE copies in the analysed genomes. A second pipeline automatically constructs phylogenetic trees using the *Caulimoviridae* EVE sequences of a given plant species, allowing the diversity of *Caulimoviridae* EVEs to be assessed for each host species, regardless of the copy number of these EVEs.

Caulifinder is a versatile tool for either producing an annotation of *Caulimoviridae* EVEs in genomes and for collecting fossil sequences that can be used to conduct evolutionary studies of *Caulimoviridae* using paleovirology approaches. It is distributed in a DOCKER image to make it easier to use. Banks are associated with it and are available in DATAInrae.

We will present the design and evaluation process of Caulifinder, carried out on sequence data from several plant genomes, and discuss the generic scope of our work for studying the evolution of viruses over long time steps.

References

[1] Patel M R, Emerman M and Malik H S 2011 Paleovirology—ghosts and gifts of viruses past *Curr. Opin. Virol.* **1** 304–9

[2] Aiewsakun P and Katzourakis A 2015 Endogenous viruses: Connecting recent and ancient viral evolution *Virology* **479–480** 26–37

[3] Teycheney P-Y, Geering A D W, Dasgupta I, Hull R, Kreuze J F, Lockhart B, Muller E, Olszewski N, Pappu H, Pooggin M M, Richert-Pöggeler K R, Schoelz J E, Seal S, Stavolone L, Umber M and Report Consortium I 2020 ICTV Virus Taxonomy Profile: Caulimoviridae *J. Gen. Virol.* **101** 1025–6

[4] Krupovic M, Blomberg J, Coffin J M, Dasgupta I, Fan H, Geering A D, Gifford R, Harrach B, Hull R, Johnson W, Kreuze J F, Lindemann D, Llorens C, Lockhart B, Mayer J, Muller E, Olszewski N E, Pappu H R, Pooggin M M, Richert-Pöggeler K R, Sabanadzovic S, Sanfaçon H, Schoelz J E, Seal S, Stavolone L, Stoye J P, Teycheney P-Y, Tristem M, Koonin E V and Kuhn J H 2018 Ortervirales: New Virus Order Unifying Five Families of Reverse-Transcribing Viruses *J. Virol.* **92** e00515-18

[5] Flutre T, Duprat E, Feuillet C and Quesneville H 2011 Considering Transposable Element Diversification in De Novo Annotation Approaches ed Y Xu *PLoS ONE* **6** e16526

[6] Quesneville H, Bergman C M, Andrieu O, Autard D, Nouaud D, Ashburner M and Anxolabehere D 2005 Combined Evidence Annotation of Transposable Elements in Genome Sequences *PLoS Comput. Biol.* **1** e22

Screening of the natural two-component systems repertoire to establish guidelines for the construction of synthetic chimeras

Emilie Cottard[1], Gwenaëlle André[1], Pierre Nicolas[1], and Sylvain Marthey[1]

[1] Université Paris-Saclay, INRAE, MaIAGE, 78350, Jouy-en-Josas, France

Corresponding Author: emilie.cottard@inrae.fr

Abstract: Two-component systems (TCSs) are central to the interaction of prokaryotes with their environment. These signal transduction pathways typically involve a sensor histidine kinase (HK) protein, which phosphorylates a response regulator (RR) upon an external stimulus. In its phosphorylated form the RR binds to a specific DNA motif and thereby modulates the expression of certain genes. Based on sequence similarity, dozens of TCSs, presumably controlling the expression of distinct regulons, can be identified in a typical bacterial genome sequence. However the input stimulus and the regulons remain unknown for a vast majority of these systems. Engineering functional HK associating the sensor domain (and hence the input stimulus) of one HK to the kinase domain of another HK with a well characterized RR might help to address this knowledge gap and is very interesting in a synthetic biology perspective. To understand how artificial HK chimeras could be built, we aim to identify natural chimeras that already exist in prokaryotic genomes. For this purpose, we developed a computational workflow which performs all pairwise local alignments within a set of representative HKs. Since the origin of chimeras may be very ancient and their sequences may have diverged significantly, we chose to use the HH-suite3 tool [1] which takes into account information on the conservation of the secondary structure in addition to the conservation of functional domains. Preliminary results obtained on a set of 148,583 HK proteins retrieved from the P2CS database [2] are currently being analyzed and will be shortly discussed.

Keywords: Two component-system, sensor histidine kinase, chimera, synthetic biology.

References

[1] Steinegger M, Meier M, Mirdita M, Vöhringer H, Haunsberger S J, and Söding J (2019) HH-suite3 for fast remote homology detection and deep protein annotation, BMC Bioinformatics, 473. doi: 10.1186/s12859-019-3019-7

[2] Ortet P, Whitworth DE, Santaella C, Achouak W, Barakat M. P2CS: updates of the prokaryotic two-component systems database. Nucleic Acids Res. 2015 Jan;43(Database issue):D536-41. doi: 10.1093/nar/gku968. Epub 2014 Oct 16. PMID: 25324303; PMCID: PMC4384028.

Extensive Characterisation of Mitochondrial Genomes in Chemically Induced Mouse Liver Tumours

Maëlle Daunesse[1], Sarah J. Aitken[2,3], Maša Roller[1], Núria López-Bigas[4], Colin A. Semple[5], Duncan T. Odom[3], Martin S. Taylor[5] and Paul Flicek[1,3]

[1] European Molecular Biology Laboratory, European Bioinformatics Institute, Hinxton, UK
[2] MRC Toxicology Unit, University of Cambridge, Cambridge, UK
[3] Cancer Research UK, Cambridge Institute, Cambridge, UK
[4] Institute for Research in Biomedicine, Barcelona, Spain
[5] MRC Institute of Genetics and Molecular Medicine, MRC Human Genetics Unit, Edinburgh, UK,
[6] Division Regulatory Genomics and Cancer Evolution, DKFZ, Heidelberg, Germany

Corresponding author: daunesse@ebi.ac.uk

Hepatocellular carcinoma (HCC) is the most common type of primary liver cancer. Multiple driver gene mutations and oncogenic pathways involved in its development have been identified, but the role that mitochondrial DNA (mtDNA) may play in it has been less investigated. Studies on mtDNA have been limited by small numbers of individuals or large genetic and environmental variability. To overcome these limitations, we leveraged tumour whole-genome sequencing and RNA-seq experiments to comprehensively characterise mitochondrial genomes in a diethylnitrosamine (DEN)-induced carcinogenesis model [1] across mouse strains with distinct cancer susceptibility and progression.

We devised a novel heteroplasmy detection approach that accounts for the circular nature of the mitochondrial genome and efficiently filters out false-positive heteroplasmies caused by nuclear mitochondrial read (NUMT) misalignment. Using this method on 760 tumour genomes, we performed heteroplasmy detection, mtDNA gene content and expression analysis in four mouse strains: Mus musculus C3H/HeOuJ, Mus musculus C57BL6/J, Mus musculus castaneus CAST/EiJ, and Mus caroli CAROLI/EiJ. Despite their recent divergence, the mouse strains investigated had different mtDNA mutation burdens, per-gene mutation rates and mtRNA expression. However, the forces shaping their mutational signatures were similar: most mutations were the product of replication-coupled DNA damage and there was a neutral selective pressure for missense and loss of function mutations. Finally, tumours had a lower mtDNA content than normal controls and the content was negatively correlated with tumour stage, suggesting a role for mtDNA in tumour progression.

Our results provide insights into mitochondrial heteroplasmy and expression in tumour development across distinct mouse genomes and its association with cancer susceptibility, transformation timing, and driver gene choice.

References

[1] Frances Connor, Tim F. Rayner, Sarah J. Aitken, Christine Feig, Margus Lukk, Javier Santoyo-Lopez, and Duncan T. Odom. Mutational landscape of a chemically-induced mouse model of liver cancer. 69(4):840–850.

A comprehensive map of preferentially located motifs reveals novel proximal *cis*-regulatory elements in plants

Julien ROZIERE [1,2,3], Cécile GUICHARD [1,2], Véronique BRUNAUD [1,2], Marie-Laure MARTIN [1,2,4] and Sylvie COURSOL [3]

[1] Université Paris-Saclay, CNRS, INRAE, Univ Evry, Institute of Plant Sciences Paris Saclay (IPS2), 91405, Orsay, France

[2] Université Paris Cité, CNRS, INRAE, Institute of Plant Sciences Paris Saclay (IPS2), 91405, Orsay, France

[3] Université Paris-Saclay, INRAE, AgroParisTech, Institut Jean-Pierre Bourgin (IJPB), 78000, Versailles, France

[4] Université Paris-Saclay, INRAE, AgroParisTech, UMR MIA-Paris, 75005, Paris, France

Corresponding Author: julien.roziere@inrae.fr

The identification of *cis*-regulatory elements controlling gene expression is an arduous challenge that is being actively explored to discover the key genetic factors responsible for traits of agronomic interest. Here, we have used a *de novo* and genome-wide approach for preferentially located motif (PLM) detection to investigate the proximal *cis*-regulatory landscape of *Arabidopsis thaliana* and *Zea mays* [1]. We report three groups of PLMs in each gene-proximal region and emphasize conserved PLMs in both species, particularly in the 3'-gene-proximal region. Comparison with resources of transcription factor and microRNA binding sites indicates that 79% of the identified PLMs are unassigned, although some are supported by MNase-defined cistrome occupancy analysis [2]. Enrichment analyses further reveal that unassigned PLMs provide functional predictions distinct from those inferred by transcription factor and microRNA binding sites. Our study provides a comprehensive map of PLMs and points at their potential utility for future characterization of orphan genes in plants.

Acknowledgements

This work was supported by the Plant2Pro® Carnot Institute in the frame of the PLMViewer program. Plant2Pro® is supported by ANR (agreement #19 CARN 0024 01). The IPS2 and IJPB laboratories benefit from the support of Saclay Plant Sciences-SPS (ANR-17-EUR-0007). J.R. is supported by a PhD fellowship of the Doctoral School 'Structure and Dynamics of Living Systems' from the University Paris-Saclay.

References

1. Rozière J, Guichard C, Brunaud V, Martin M-L, Coursol S. Genome-wide identification of preferentially located motifs in gene-proximal regions provides new insights into cis-regulatory sequences in plants. bioRxiv 2022.01.17.476590; doi: https://doi.org/10.1101/2022.01.17.476590
2. Savadel SD, Hartwig T, Turpin ZM, Vera DL, Lung PY, Sui X, Blank M, Frommer WB, Dennis JH, Zhang J, Bass HW. The native cistrome and sequence motif families of the maize ear. PLoS Genet. 2021 Aug 12;17(8):e1009689. doi: 10.1371/journal.pgen.1009689. PMID: 34383745; PMCID: PMC8360572.

Long reads RNA sequencing analysis with Oxford Nanopore Technologies: Comparison of different library protocols and bioinformatics processing

Asmae Bachr[1], Aurélie Leduc[1], Céline Derbois[1], Marc Delépine[1], Florian Sandron[1], Éric Cabannes[1], Jean-François Deleuze[1] and Vincent Meyer[1]

[1] CNRGH, 2 Rue Gaston Crémieux, 91000 Evry, France

Corresponding author: asmae.bachr@cnrgh.fr

Recently, long fragment sequencing technologies have been introduced into the field of RNAseq, offering the possibility to directly generate reads that can cover the entire length of the transcript. This third-generation technology is characterized by real time sequencing of molecules enabling the generation of long reads that can reach several hundred kilobases. These long reads allow a better identification of alternatively spliced forms and new transcripts.

We studied the impact of different ONT library preparations on the main informative quality metrics such as the yield, the read size distribution, the transcript coverage and 3' bias related to the sample preparation, the transcript quantification and identification. With this evolving technology, we also performed benchmarks of different bioinformatics tools.

The human GM12878 cell line was used to compare three library protocols including direct RNA, cDNA-PCR total and cDNA-PCR polyA+. Each library was spiked with a low amount (1%) of SIRVs (Spike-in RNA variant) corresponding to a group of synthetic transcripts that mimic transcriptome complexity. We used these spike-in data to evaluate the performance of different tools (StringTie2 [1], Salmon [2] and FLAIR [3]) for transcripts quantification and characterization. As good performance on SIRV data was obtained, we decided to test them on human data to make the protocol comparison.

Firstly, we compared the different informative quality metrics mentioned previously. Results show cDNA-PCR total and poly A+ have the best throughput compared to the direct RNA. Nevertheless, direct RNA-seq provides the best mean read size and is less impacted by 3' bias as expected.

After this primary analysis, we looked at human transcripts. On the one hand, we focused our analysis on known human transcripts only (Ensembl annotation release 92) to understand the coverage profile (depth and breadth). We noticed transcript coverage variations coming from full-length reads ratio related to each experimental condition. The direct RNA sequencing shows a lower number of identified transcripts and a better coverage compared to the other cDNA-PCR total and polyA+ library protocols.

On the other hand, we tried to directly reconstruct human transcripts from « full-length » reads. We tested the following tools: StringTie2 based on an assembly approach and FLAIR based on a mapping approach. GffCompare tool was then used to evaluate the accuracy of identified transcripts compared to reference transcripts provided in the Ensembl annotation file. The comparison between these two strategies shows major differences regarding the number and type of detected transcripts.

This study shows that ONT technology allows full-length coverage of transcripts. Nevertheless, the library preparation and sequencing will lead to biases and limitations for the analysis. Frequently evolving bioinformatics pipelines provide very different results that will require filtrations to limit the impact of transcript quantification and identification analyses.

References

1. Kovaka S., Zimin A.V., Pertea G.M., Razaghi R., Salzberg S.L., Pertea M. Transcriptome assembly from long-read RNA-seq alignments with StringTie2. Genome Biol, vol. 20, n° 278, 2019.

2. Patro R., Duggal G., Love M.I., Irizarry R.A., and Kingsford C. Salmon provides fast and bias-aware quantification of transcript expression. Nat Methods, 14, 2017.

3. Tang A.D., Soulette C.M., van Baren M.J., Hart K., Hrabeta-Robinson E., Wu C.J., et al. Full-length transcript characterization of SF3B1 mutation in chronic lymphocytic leukemia reveals downregulation of retained introns. Nat Commun, 11, 2020.

MYC deficiency impairs the development of effector/memory T lymphocytes

Mathis Nozais[1,3], Marie Loosveld[1,2,3], Saran Pankaew[1], Clemence Grosjean[1], Noemie Gentil[1], Julie Quessada[1], Bertrand Nadel[1], Cyrille Mionnet[1], Delphine Potier1[1], and Dominique Payet-Bornet[1]

[1] Aix Marseille Univ, CNRS, INSERM, Centre d'Immunologie de Marseille-Luminy (CIML)
[2] APHM, Hopital La Timone, Laboratoire d'Hématologie, Marseille, France
[3] These authors contributed equally

Corresponding Author: payet@ciml.univ-mrs.fr & potier@ciml.univ-mrs.fr

In the thymus, T cell progenitors differentiate in order to generate naïve T lymphocytes which migrate to secondary lymphoid organs where they can encounter an antigen. Antigenic stimulation of naive T cells induces their proliferation and differentiation into effector and memory T cells (Eff/mem T cells). During thymic differentiation, genomic alterations in thymocytes can promote the development of T-cell acute lymphoblastic leukemia (T-ALL)[1]. One of the most potent alterations is the loss-of-function mutation of *PTEN* gene. PTEN is a tumor suppressor, which, in T-ALL, can sustain the oncogenic activation of MYC[2]. The well-known *MYC* oncogene[3] is rarely mutated in T-ALL, nevertheless, it appears to be a key factor in T-ALL leukemogenesis. In line with human data, inactivation of *Pten* gene during murine thymopoeisis ($Pten^{del}$ mouse model) gives rise to T-ALL that systematically express high level of MYC protein[4].

Herein our objective was to investigate the impact of MYC inactivation on physio-pathological development of T cells. To do so we used mouse models in which *Myc* and/or *Pten* genes were deleted during thymopoeisis. Thus, we analysed four types of mice: Control; $Pten^{del}$; Myc^{del}; and double knock-out, $Myc^{del}Pten^{del}$. These mice were also crossed with eYFP reporter mice, which allow the tracking of cells that have inactivated *Myc* and/or *Pten*. First, we observed that $Myc^{del}Pten^{del}$ mice do not develop tumors. This finding confirms that MYC expression is absolutely required for PTEN loss-mediated tumor transformation. Then, thymus and spleens from the different genotypes were analyzed by multiplexed single-cell RNA sequencing (scRNAseq) approaches[5]. After data pre-processing and sample demultiplexing, we assigned cell type to the 23 identified clusters according to various gene markers. We notably distinguished the main T cell subsets of thymus and spleen. Our data show that *Myc* deletion at the CD4+CD8+ stage does not affect terminal differentiation of thymocytes, while it disrupts splenic T cell homeostasis. Indeed, thanks to eYFP reporter gene, we observed a strong diminution of Eff/mem T cells in Myc^{del} and $Myc^{del}Pten^{del}$ mice. Using Differential Gene Expression (DGE) and Gene Ontology, we found that MYC-deficient naive T cells down-regulate genes related to ribosomes and protein synthesis, indicating that cell proliferation is disturbed. Moreover, our scRNAseq analysis reveals that the number of a small T-cell subset ($TCR\gamma\delta^+$ T-cells) increases in *Myc*-deficient spleen, such expansion appears to be a collateral effect of MYC inactivation in thymocytes. *In silico* analysis were reinforced by biological experiments. Altogether our result[6] show that in absence of MYC, naïve T cells can be activated by an antigen, nevertheless the proliferation and differentiation of these MYC-deficient cells into Eff/mem T cells are inhibited.

References

[1] Girardi T, Vicente C, *et al.* The genetics and molecular biology of T-ALL. Blood. 2017;129(9):1113–1123. doi:10.1182/blood-2016-10-706465

[2] Bonnet M, Loosveld M, *et al.* Posttranscriptional deregulation of MYC via PTEN constitutes a major alternative pathway of MYC activation in T-cell acute lymphoblastic leukemia. Blood. 2011;117(24):6650–6659. doi:10.1182/blood-2011-02-336842

[3] Dang CV. MYC, metabolism, cell growth, and tumorigenesis. Cold Spring Harbor Perspectives in Medicine. 2013;3(8). doi:10.1101/cshperspect.a014217

[4] Gon S, Loosveld M, *et al.* Fit αβ T-cell receptor suppresses leukemogenesis of Pten-deficient thymocytes. Haematologica. 2018;103(6):999–1007. doi:10.3324/haematol.2018.188359

[5] Stoeckius M, Zheng S, *et al.* Cell Hashing with barcoded antibodies enables multiplexing and doublet detection for single cell genomics. Genome Biology. 2018;19(1):224. doi:10.1186/s13059-018-1603-1

[6] Nozais M, Loosveld M, *et al.* MYC deficiency impairs the development of effector/memory T lymphocytes. iScience. 2021;24(7):102761. doi:10.1016/j.isci.2021.102761

Analysis of single-cell RNA-seq human PBMC datasets

Marie-Ange PALOMARES[1], Céline DERBOIS[1], Jean-François DELEUZE[1], Eric CABANNES[1] and Eric BONNET[1]

[1] Centre National de Recherche en Génomique Humaine (CNRGH), Institut de Biologie François Jacob, CEA, Université Paris-Saclay, Evry, France.

Corresponding Author: eric.bonnet@cnrgh.fr

DNA sequencing technology has scaled up very rapidly in throughput but also scaled down in terms of the amount of DNA that is required for analysis, to the point that it is now commonplace to analyze the DNA and RNA content of individual cells. Single-cell approaches have triggered previously impossible applications in basic research and clinical science, like transcriptome analysis of rare circulating tumor cells, characterization of early differentiation events in human embryogenesis, investigation of cell fate choices and creation of large-scale cell atlases such as the human cell atlas.

However, integrated analysis of different scRNA-seq data sets, consisting of multiple transcriptomic subpopulations or integrating measurements produced by different technologies, remains challenging. It is especially difficult to distinguish between the composition of cell types in a sample and expression changes within a given cell type. Furthermore, sample preparation is a quite crucial step for single-cell analysis, with many possible options of preparation to obtain the good quality cells that correctly reflect the biological conditions in the original tissue.

In this study, we used single-cell RNA-seq (scRNA-seq) to assess the expression of peripheral blood mononuclear cells (PBMC) from six different samples. PBMC were isolated from a healthy anonymous donor's whole blood specimen provided by EFS (Etablissement Français du Sang), using a Ficoll gradient. PBMC were then splitted into 6 different samples with different treatments. One sample was set to rest (resting), one sample was stimulated with LPS (incubation for 4 hours with 1µg/mL LPS in order to induce an inflammatory response), 2 samples were processed directly and 2 samples were first frozen (according to the 10X Genomics protocol for frozen samples), then 8 days later were thawed and processed.

Samples were prepared with the Chromium Next GEM Single Cell 3' GEM v3.1 kit following the manufacturer recommandations (10X Genomics). Libraries were sequenced on a Novaseq 6000 sequencer (Illumina). Sample demultipexing, barcode processing and unique molecular identifiers (UMI) counting were performed by using the 10X Genomics pipeline Cellranger v6.0.1 with default parameters. Analyses were performed using the software tools Seurat, Scrublet and MultiMAP.

For this study, we used the resting versus stimulated data to perform and compare three data integration techniques on our dataset. Then we compared the fresh versus frozen samples to analyze the effect of congelation on the quality of the results.

The resting and stimulated datasets were sequenced with a target of 8000 cells/sample. After processing with CellRanger, we obtained a total of 14,771 viable cells for the analysis, with 114,000 reads/cell and a sequencing saturation above 76%. After removing low quality cells, we performed normalization, scaling and dimension reduction (UMAP), where we could visualize a clear difference between resting and stimulated groups of cells. We then applied three different algorithms to integrate the data: canonical correlation analysis (CCA), reciprocal PCA (RPCA) and MultiMAP (MMAP). The three different techniques effectively perform integration, but we can visually see differences in the results.

For the fresh/frozen samples, we aimed at 6000 cells/sample, obtaining after CellRanger processing a total of 16,018 cells with a mean of 111,000 reads/cell and an average sequencing saturation of 81% per sample. After quality control and removal of low quality cells, we compared the number of genes per cell, the number of molecules per cell and the percentage of mitochondrial genes per cell. We also visualized the cells after UMAP dimensional reduction. We can clearly see many differences between the fresh and frozen samples, despite the fact that they are all technical replicates originating from the same initial pool of cells. We conclude that the congelation step is clearly lowering the quality of the results for PBMC single-cell RNA-seq analysis.

Impact of genomic variation on CTCF binding and 3D genome organization in breast cancer cells

Julie Segueni[1], Joanne Edouard[1], Daan Noordermeer[1]

[1] Université Paris-Saclay, CEA, CNRS, Institute for Integrative Biology of the Cell (I2BC), Gif-Sur-Yvette, France

Corresponding Authors: julie.segueni@i2bc.paris-saclay.fr, daan.noordermeer@i2bc.paris-saclay.fr

Mammalian genomes are organized into several thousands of Topologically Associating Domains (TADs) where intra-domain 3D interactions are strongly favored over neighboring regions. TADs constitute regulatory modules where interactions between enhancers and promoters occur preferentially inside the TAD. In contrast, the binding of the CTCF insulator protein at the boundaries of TADs prevents the formation of promoter-enhancer loops between neighboring domains.

Several recent studies have reported that CTCF binding sites (CBS) are mutation hotspots in cancer cells. Moreover, the perturbation of CTCF binding at TAD boundaries can cause 'enhancer hijacking', whereby oncogenes are activated through the formation of loops with enhancers in neighboring TADs [1]. Until now, the effect of genome structural variation to oncogene activation within the context of reorganized CTCF binding and TAD restructuration has not been reported.

In my project, we aim to identify mutations within CTCF binding sites, followed by the characterization of changes in CTCF binding, reorganization of TADs, changes in promoter-enhancer loops and changes in gene expression. For this purpose, I am integrating newly generated genomics data from two widely-used breast cancer cell lines (MCF7 and T47D) with data from a control breast epithelial cell line (MCF10A).

To obtain both qualitative (absence or presence) and quantitative (increased or decreased) insights into changes in CTCF binding, I used an analysis strategy based on input and spike-in normalization of ChIP-seq data. A comparison of the breast cancer cell lines to the control cell line identified an unexpectedly large number of differentially bound peaks: 5% of peaks were identified as significantly different, with around 1/3 of peaks showing a complete gain or loss of binding. Among these differential peaks, many peaks were shared between cell lines.

To assess the impact of differential CBS on TAD boundaries and loops, I analyzed high-resolution Hi-C data, including normalization for Copy Number Variations in each cell line. Integration of our identified TAD boundaries with differences in CBS and gene expression (both newly-generated RNA-seq and 'gold standard' microarray data [2]) has identified large numbers of deregulated genes within reorganized TADs. Exploration of sequence determinants at differential CBS (structural variation, DNA methylation, TF binding) is currently ongoing. Importantly, CTCF binding is known to be methylation-sensitive [3]. Furthermore, genome and epi-genome editing experiments will be used to validate the impact of CBS perturbations on 3D genome organization and gene expression in our breast cancer cell model.

References

1. D. Hnisz *et al.*, « Activation of proto-oncogenes by disruption of chromosome neighborhoods. », *Science*, vol. 351, nº 6280, p. 1454-1458, mars 2016, doi: 10.1126/science.aad9024.
2. R. M. Neve *et al.*, « A collection of breast cancer cell lines for the study of functionally distinct cancer subtypes. », *Cancer Cell*, vol. 10, nº 6, p. 515-527, déc. 2006, doi: 10.1016/j.ccr.2006.10.008.
3. H. Wang *et al.*, « Widespread plasticity in CTCF occupancy linked to DNA methylation. », *Genome Res.*, vol. 22, nº 9, p. 1680-1688, sept. 2012, doi: 10.1101/gr.136101.111.

GenomiqueENS, the IBENS Genomics core facility

Corinne Blugeon[1], Ali Hamraoui[1], Laurent Jourdren[1], Sophie Lemoine[1], Catherine Senamaud-Beaufort[1], Stéphane Le Crom[2,1], and Morgane Thomas-Chollier[1]

[1] GenomiqueENS, Institut de Biologie de l'ENS (IBENS), Département de biologie, École normale supérieure, CNRS, INSERM, Université PSL, 75005 Paris, France

[2] Sorbonne Université, CNRS, Institut de Biologie Paris-Seine (IBPS), Laboratory of Computational and Quantitative Biology (LCQB), F-75005, Paris

Corresponding Author: jourdren@bio.ens.psl.eu

The **genomics core facility of the Institut de Biologie de l'École normale supérieure** (IBENS) [1,2] was created in 1999. We focus on functional genomics in **eukaryotes** , including classical model organisms, as well as more exotic organisms (jellyfish, birds, butterflies...). **The facility has always been a well-balanced structure between wet-lab and bioinformatics**: half of the team is involved on the wet-lab part; the other half being involved on the data analysis part. Our goal is to assist laboratories during their **high-throughput sequencing projects** from the experimental design to data analysis for publication. We are part of the **France Génomique consortium** and have been certified with the **ISO 9001** quality international standard since March 2013.

Our genomics core facility offers many services to the community: **library preparation** (RNA-seq, scRNA-seq and ChIP-seq), **sequencing** (including "ready-to-load" libraries) for short (Illumina) and long reads (Oxford Nanopore); and **bioinformatics analysis** (RNA-seq and scRNA-seq).

All the staff working on the facility gets a balanced schedule between the core **production service** and **research and development projects** to propose **up-to-date and reliable experimental solutions** to our collaborators. To cope with the experimental constraints of our users among the research teams, we invest time in **testing library protocols** (very low quantities, ribosome depletions...). We are also deeply involved in **software development** to manage our project analyses (65% of projects are analysed on the facility). The tools we develop are distributed on an open source basis on **GitHub** [3] and we now provide most of them as **Docker** images [4] to **ease the distribution** of our work. We develop workflows to achieve reproducible and transparent data analysis of our high throughput experiments.

Since 2016, our facility has been offering two new technologies. The first one is devoted to **single cell RNA-seq** with a Chromium system from **10X Genomics** based on the Drop-seq protocol. The second one is dedicated to long read sequencing in RNA-seq. We use **Oxford Nanopore Technologies sequencers** devices in order to sequence full length transcripts for isoform abundance estimation.

Over the last few months we have released a rewritten and enhanced version of **ToulligQC** [5], our QC tool for Oxford Nanopore sequencers and we are currently testing **scNaUmi-seq** protocol [6] to propose our scRNA-seq service combined with long read sequencing.

All these developments allow us to be at the state of the **art in functional genomics applications**, so that we can provide to our users all the tools needed to succeed in their high throughput experiments.

Acknowledgements

The IBENS genomics core facility 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] https://genomique.bio.ens.psl.eu/
[2] Twitter @Genomique_ENS
[3] https://github.com/GenomiqueENS
[4] https://hub.docker.com/r/genomicpariscentre/
[5] https://github.com/GenomiqueENS/toulligQC
[6] Lebrigand, K., *et al.* High throughput error corrected Nanopore single cell transcriptome sequencing. *Nat Commun* **11**, 4025 (2020)

Research and development at the I2BC Next-generation Sequencing Facility: an overview

Kévin GORRICHON[1*], Delphine NAQUIN[1*], Rania OUAZAHROU[1], Yan JASZCZYSZYN[1], Claude THERMES[1], Erwin VAN DIJK[1] and Céline HERNANDEZ[1]

[1] Université Paris-Saclay, CEA, CNRS, Institute for Integrative Biology of the Cell (I2BC), 91198, Gif-sur-Yvette, France

Corresponding Author: I2BC-sequencage@i2bc.paris-saclay.fr

Core facilities aim to give their users access to the newest technologies and scientific methods, which appear and evolve rapidly. Since its creation in 2010, the mission of the I2BC next-generation sequencing facility (PSI2BC) is to provide the scientific community, whether academic or industrial, with services and support in the domain of high throughput sequencing and its applications in functional genomics and transcriptomics. We present an overview of recent research and development (R&D) projects carried out at our core facility, and based on both Illumina and Oxford Nanopore Technologies (ONT).

Firstly, we present our developing expertise in the study of DNA and RNA modifications using ONT sequencing:

- detection of 5-methylcytosine (5mC) modifications of DNA, a well-characterized mark associated with transcriptional repression.
- targeted ONT sequencing of specific genomic regions using a Cas9 directed approach. We have tested a novel ONT kit based on a technology similar to a recently published method (nCATS), which allowed us to strongly enrich for a specific region of the mouse genome.
- detection of pseudo-uridine, the most widespread modification in RNA, present in all living organisms.

We also mention other advances including:

- generation of ultra-long ONT reads. We have tested a recently released ONT kit and have obtained a large proportion of reads of tens to hundreds of kilobases in length.
- development of an improved small RNA Illumina library preparation method with less bias and better detection of 2'-O-Methyl RNAs [1]. Several types of RNA such as plant microRNAs (miRNAs) carry a 2'-O-Me modification at their 3' terminal nucleotide. This modification complicates library preparation as it inhibits 3' adapter ligation. Our protocol has less overall bias and is less affected by the modification than standard methods.

Acknowledgements

The I2BC sequencing facility is supported by France Génomique (funded by the French National Program "Investissement d'Avenir" ANR-10-INBS-09).

References

1. Erwin L. van Dijk and Claude Thermes. *A Small Rna-Seq Protocol with Less Bias and Improved Capture of 2′-o-Methyl RNAs*. In RNA Modifications, Mary McMahon, 2298:153-67. New York, NY: Springer US, 2021.

Predicting clinical response to immunotherapy in advanced melanoma

Matthieu GENAIS[1], Anne MONTFORT[1], Bruno SÉGUI[1] and Vera PANCALDI[1]
INSERM Centre de Recherche en Cancerologie de Toulouse,2 Avenue Hubert Curien, 31100 Toulouse, FRANCE

Corresponding author: matthieu.genais@inserm.fr

1 Introduction

Immune checkpoint inhibitors (ICI) such as anti-PD-1 act on T cells to restore their ability to kill cancer cells. Cutaneous melanoma is a poor-prognosis skin cancer that can be treated by ICI. Despite major advances in the field of immunotherapy, melanoma kills half of all patients within 5 years of treatment induction due to primary or acquired resistance. Over the last 10 years, our team has identified a mechanism of resistance to immunotherapy that depends on the production of TNF, a major inflammatory cytokine that acts as a brake on the immune response against tumours in mouse melanoma models [1][2] and could ameliorates adverse events due to immunotherapy treatment [3].

2 Materials and method/Results

Bulk RNA-seq on tumour samples from TNF knock-out (KO) mice and wild type mice were analysed by standard differential gene expression analysis [4] and pathway enrichment analysis [5] , as well as by deconvolution [6] [7] of cell types and inference of transcription factor activities [8] [9]. We showed that knocking out TNF-alpha impacts the immune response in multiple ways, for example by up-regulating the humoral immune response or by down-regulating mitochondrial pathways. Moreover, knocking out TNF seems to have an inconsistent impact on mice. Indeed, using single sample GSEA (Gene Set Enrichment Analysis [10]), we discovered 2 groups of TNF KO mice: one that displays a phenotype close to anti-PD1 treated mice, and one with a phenotype that is close to wild type mice. These groups can be retrieved by computing differential transcription factor (TF) activities between TNF KO mice and WT mice with a list of up-activated and down-activated TFs. TF activities of the first group of mice correlated positively with immune scores and infiltrated deconvolved immune cells (CD8) and the opposite for the second one. We then classified mice samples into hot and cold tumours, respectively.

Finally, using this TNF KO TF activity signature in human melanoma patients (pre-treatment to anti-PD1 therapy), we distinguished the same two profiles that we had seen in mice by creating a TF activities based score (TF score). We saw an enrichment of responders to anti-PD1 treatment for high score patients compared to low score patients. High score patients in the pre-treatment condition seem to exhibit a similar profile to TNF KO mice, suggesting that TNF might not be essential for hot tumour patterns in patients who are the most susceptible to respond to immunotherapy, as observed in mice.

3 Conclusion

Our RNA-seq data from tumour samples in mice allowed us to discover a set of TFs that become activated in some TNF KO mice. This TF set could also be found in anti-PD1 or TNFKO+anti-PD1. We used these TF along immune scores and deconvolution data to distinguish between hot (highly infiltrated/high immune scores) and cold tumours (lowly infiltrated/low immune scores). Instead of looking directly at gene expression, we used VIPER inferred TF activities to create a score based on "TNF KO" TF score that allowed us to identify two groups of patients in public RNAseq datasets of patients treated with anti-PD1 agents. This "TNF KO" TF activity was applied to melanoma patients on pre-treatment samples, where a higher score was measured for good prognosis patients. Notably, the low activity patients could either respond or not to anti-PD1, suggesting that other mechanisms contribute to response to immunotherapy, independently of the "TNF KO" TF activity. We are currently better characterizing this TF activity signature in terms of pathways and exploring how the cold tumours (samples with low TF score) can still respond to monotherapy. We also want to test our signature on bitherapy cohorts (anti-PD1+anti-CTLA4) [11,12] and finally on our clinical trial TICIMEL wich is based on tritherapy (anti-PD1+anti-CTLA4+anti-TNF) [13].

References

[1] Florie Bertrand, Anne Montfort, Elie Marcheteau, Caroline Imbert, Julia Gilhodes, Thomas Filleron, Philippe Rochaix, Nathalie Andrieu-Abadie, Thierry Levade, Nicolas Meyer, Céline Colacios, and Bruno Ségui. TNF blockade overcomes resistance to anti-PD-1 in experimental melanoma. 8(1):2256. Number: 1 Publisher: Nature Publishing Group.

[2] Florie Bertrand, Julia Rochotte, Céline Colacios, Anne Montfort, Anne-Françoise Tilkin-Mariamé, Christian Touriol, Philippe Rochaix, Isabelle Lajoie-Mazenc, Nathalie Andrieu-Abadie, Thierry Levade, Hervé Benoist, and Bruno Ségui. Blocking tumor necrosis factor enhances CD8 t-cell-dependent immunity in experimental melanoma. 75(13):2619–2628.

[3] Elisabeth Perez-Ruiz, Luna Minute, Itziar Otano, Maite Alvarez, Maria Carmen Ochoa, Virginia Belsue, Carlos de Andrea, Maria Esperanza Rodriguez-Ruiz, Jose Luis Perez-Gracia, Ivan Marquez-Rodas, Casilda Llacer, Martina Alvarez, Vanesa de Luque, Carmen Molina, Alvaro Teijeira, Pedro Berraondo, and Ignacio Melero. Prophylactic TNF blockade uncouples efficacy and toxicity in dual CTLA-4 and PD-1 immunotherapy. *Nature*, 569(7756):428–432, May 2019.

[4] Michael I. Love, Wolfgang Huber, and Simon Anders. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. 15(12):550.

[5] Alexey A. Sergushichev. An algorithm for fast preranked gene set enrichment analysis using cumulative statistic calculation. Section: New Results Type: article.

[6] Francisco Avila Cobos, José Alquicira-Hernandez, Joseph E. Powell, Pieter Mestdagh, and Katleen De Preter. Benchmarking of cell type deconvolution pipelines for transcriptomics data. 11(1):5650. Number: 1 Publisher: Nature Publishing Group.

[7] Ting Xie, Julien Pernet, Nina Verstraete, Miguel Madrid-Mencía, Mei-Shiue Kuo, Alexis Hucteau, Alexis Coullomb, Jacobo Solórzano, Olivier Delfour, Francisco Cruzalegui, and Vera Pancaldi. GEM-DeCan: Improved tumor immune microenvironment profiling through novel gene expression and DNA methylation signatures predicts immunotherapy response. Section: New Results Type: article.

[8] Adam A. Margolin, Ilya Nemenman, Katia Basso, Chris Wiggins, Gustavo Stolovitzky, Riccardo Dalla Favera, and Andrea Califano. ARACNE: An algorithm for the reconstruction of gene regulatory networks in a mammalian cellular context. 7(1):S7.

[9] Mariano J. Alvarez, Yao Shen, Federico M. Giorgi, Alexander Lachmann, B. Belinda Ding, B. Hilda Ye, and Andrea Califano. Network-based inference of protein activity helps functionalize the genetic landscape of cancer. 48(8):838–847.

[10] Sonja Hänzelmann, Robert Castelo, and Justin Guinney. GSVA: gene set variation analysis for microarray and RNA-seq data. 14(1):7.

[11] Frank Stephen Hodi, Vanna Chiarion-Sileni, Rene Gonzalez, Jean-Jacques Grob, Piotr Rutkowski, Charles Lance Cowey, Christopher D. Lao, Dirk Schadendorf, John Wagstaff, Reinhard Dummer, Pier Francesco Ferrucci, Michael Smylie, Andrew Hill, David Hogg, Ivan Marquez-Rodas, Joel Jiang, Jasmine Rizzo, James Larkin, and Jedd D. Wolchok. Nivolumab plus ipilimumab or nivolumab alone versus ipilimumab alone in advanced melanoma (CheckMate 067): 4-year outcomes of a multicentre, randomised, phase 3 trial. 19(11):1480–1492.

[12] Jacob Schachter, Antoni Ribas, Georgina V. Long, Ana Arance, Jean-Jacques Grob, Laurent Mortier, Adil Daud, Matteo S. Carlino, Catriona McNeil, Michal Lotem, James Larkin, Paul Lorigan, Bart Neyns, Christian Blank, Teresa M. Petrella, Omid Hamid, Honghong Zhou, Scot Ebbinghaus, Nageatte Ibrahim, and Caroline Robert. Pembrolizumab versus ipilimumab for advanced melanoma: final overall survival results of a multicentre, randomised, open-label phase 3 study (KEYNOTE-006). 390(10105):1853–1862.

[13] Anne Montfort, Thomas Filleron, Mathieu Virazels, Bruno Ségui, and Nicolas Meyer. Anti-TNF+nivolumab+ipilimumab pour le traitement du mélanome avancé : premiers résultats de l'essai clinique TICIMEL. *Annales de Dermatologie et de Vénéréologie - FMC*, 1(8, Supplement 1):A121–A122, December 2021.

Automated identification of a cancer patient treatment: from sequencing to treatment prioritization

Nicolas SOIRAT[1, 2, 3], Denis Bertrand[1], Sacha BEAUMEUNIER[1], Nicolas PHILIPPE[1], Dominique VAUR[2], Sophie KRIEGER[2,3], Anne-Laure BOUGÉ[1], and Laurent CASTERA[2]

[1] SeqOne, 22 Rue Durand, 34000, Montpellier, France

[2] Laboratoire de biologie et de génétique du cancer, Centre François Baclesse, 14000, Caen, France

[3] Unicaen, Esp. de la Paix, 14000, Caen, France

Corresponding Author: nicolas.soirat@seqone.com

The emergence of sequencing allowed the scientific community to gather a tremendous amount of cancer genomic data, characterizing biomarkers responsible for tumorigenesis that might indicate potential treatments. The use of short-read sequencing to identify cancer patient treatment is becoming a more common practice in hospitals [1,2]. To standardize the treatment identification some prediction frameworks have been developed, but they mostly focus on a single alteration type and very few have been implemented.

We design a targeted DNA and RNA panel covering 639 cancer genes and 57 fusion genes to obtain a comprehensive patient genomic landscape. We developed a decisional algorithm which prioritizes all known variant-therapy associations. Several rules give a score for each association based on more than 20 variant features indicating the variant impact in cancer, the patient indication and similarity of patient variant with variant in therapeutic databases.

We generated a thousand simulated tumors, each containing passenger mutations and a targetable mutation from the Civic database. Our method correctly classifies the targetable mutation in its top predictions (average rank 2.19). Furthermore, on a cohort of 12 patients, we obtain similar results as 2 clinical routine approaches using our fully automated protocol. We are planning to expand our validation to a pan cancer cohort of 500 patients supported by therapeutic reports.
We design a complete framework for multiple variant drug association identification in order to make easier therapeutic choices for a clinician. We succeed to integrate it into our variant calling workflow and show good performance of our method to prioritize targetable variants.

References

1. AACR Project GENIE: Powering Precision Medicine through an International Consortium. *Cancer Discov.* **7**, 818–831 (2017).
2. Tamborero, D. *et al.* The Molecular Tumor Board Portal supports clinical decisions and automated reporting for precision oncology. *Nat. Cancer 2022 32* **3**, 251–261 (2022).

RiboMethSeq platform at CRCL/CLB to profile ribosomal RNA 2'O-ribose methylation

Théo COMBE[1,4], Hermes PARAQINDES[1,2], Janice KIELBASSA[1], Jessie AUCLAIR[3], Marjorie CARRERE[3], Valery ATTIGNON[3], Alain VIARI[1,5], LAURIE TONON[1,4], Emilie THOMAS[1,4], Anthony FERRARI[1,4], Virginie MARCEL[2]

[1] Bioinformatics Platform Gilles Thomas, CRCL/CLB, 28 rue laennec, 69008, Lyon, France
[2] Ribosome, Translation and Cancer team, CRCL, 28 rue Laennec, 69008, Lyon, France
[3] Cancer Genomics Platform, CRCL/CLB, 28 rue Laennec, 69008, Lyon, France
[4] Fondation Synergie Lyon Cancer, 28 rue Laennec, 69008, Lyon, France
[5] INRIA Grenoble-Rhône-Alpes, 655 Avenue de l'Europe, 38334, Montbonnot, France

Corresponding Author: Theo.COMBE@lyon.unicancer.fr

The ribosome is well-known as the cellular machinery translating the messenger RNAs (mRNAs) into proteins. It is composed of 80 proteins and 4 ribosomal RNAs (rRNA) in humans. While it was mostly believed that the ribosome displays always the same activity, recent studies have revealed that ribosomes exhibit different compositions depending on the physio-pathological contexts. At rRNAs level, variation in the chemical modifications play a role in translational control. We contributed in demonstrating that rRNA 2'O-ribose methylation (2'Ome), corresponding to the addition of a methyl group on the nucleotide's ribose 2' hydroxyl, does not only regulate the intrinsic activity of the ribosomes but is also altered at some specific rRNA sites in cancer [1,2,3].

RiboMethSeq is an innovative RNASeq-based approach that has been developed in 2015 to quantify rRNA 2'Ome at all the sites at once in yeast [4]. It is based on a partial alkaline hydrolysis. Indeed, presence of a 2'Ome prevents alkaline hydrolysis between the 2'Ome nucleotide and the following one. Since this bond is preserved in presence of 2'Ome, reads do not end at the 2'Ome position or start at the following one. Thus, comparing 5' and 3' end reads counts at each rRNA position allows to analyse rRNA 2'Ome level.

At Centre Leon Bérard and Cancer Research Center of Lyon, a RiboMethSeq platform was created in 2018, with a complete workflow from sample preparation to bioinformatic analyses. Usage of Illumina NovaSeq sequencer allows the sequencing of 48 samples in a single run, thus reducing both batch effects and costs. Sequencing data are stored on the High Performance Computing cluster of the Gilles Thomas Bioinformatic platform, involved in clinical, translational and basic research programs and composed of diverse expertises (biostaticians, bioinformaticians, data managers, developers…).

Regarding RiboMethSeq, in addition to performing routine analysis of human samples from cell lines or large collection of biopsies, we also develop novel methodological and bioinformatic tools, since dedicated QC or analytic pipelines are missing. With this poster, we will present an overview of the RiboMethSeq platform as well as the bioinformatic tools we have developed.

§ Acknowledgements

The implementation and development of this platform is founded by Institut National du Cancer (INCa, PLBio MARACAS) and SIRIC program (LyriCAN).

§ References

[1] Marcel, Virginie et al. *P53 Acts as a Safeguard of Translational Control by Regulating Fibrillarin and rRNA Methylation in Cancer*. Cancer cell. 24. 318-30. 10.1016/j.ccr.2013.08.013, 2013.

[2] Erales, Jenny et al. *Evidence for rRNA 2'-O-methylation plasticity: control of intrinsic translational capabilities of human ribosomes*. Proc. Natl Acad. Sci. USA114, 12934–12939, 2017.

[3] Jaafar, Mariam et al. *2'O-Ribose Methylation of Ribosomal RNAs: Natural Diversity in Living Organisms, Biological Processes, and Diseases*. Cells 2021, 10, 1948.

[4] Birkedal, Ulf et al. *Profiling of ribose methylations in RNA by high-throughput sequencing*. Angewandte Chemie (International ed. in English) vol. 54,2 (2015): 451-5.

Rattus norvegicus reference genome evaluation for hippocampus RNA-seq data analysis: a glimpse into spatial transcriptomics

Christophe Le Priol[1] and Andrée Delahaye-Duriez[1,2,3]
[1] Inserm UMR 1141 NeuroDiderot, Université de Paris Cité, Paris, France
[2] UFR Santé Médecine Biologie Humaine, Université Sorbonne Paris Nord, Bobigny, France
[3] Unité fonctionnelle de médecine génomique et génétique clinique, Hôpital Jean Verdier, AP-HP, Bondy, France

Corresponding author: christophe.le-priol@inserm.fr

The emergence of spatial RNA-seq enabled to preserve spatial information in histological sections while profiling transcriptomes [1]. One of the first steps of a usual RNA-seq data analysis workflow consists in quantifying gene expression by aligning the sequencing reads to a reference genome and counting the aligned reads in its annotated regions. Downstream analysis, such as the identification of differentially expressed genes, strongly rely on the quality of this process. Similarly to most of single-cell RNA-seq technologies, the refinements brought by spatial RNA-seq technologies can be mitigated by a lower sequencing depth in comparison with bulk RNA-seq.

Here, we propose to evaluate the effect of widely used *Rattus norvegicus* reference genomes consisting of Ensembl and RefSeq annotations of the Rnor_6.0 assembly and the new mRatBN7.2 assembly recently published by RefSeq [2] on a classical differential expression workflow. We re-analyzed published bulk RNA-seq datasets from different hippocampal regions [3,4]. We also tested the choice of reference genome on a newly generated spatial RNA-seq dataset (10X Visium technology on hippocampus sections in a rat model for mesial temporal lobe epilepsy).

We revealed that the annotations of the new *Rattus norvegicus* genome assembly provide an improvement of read mapping statistics. Besides, using the consistency and stringency metrics introduced by Chen *et al* [5], we highlighted discrepancies of at least 5% in read counts when using Ensembl and RefSeq annotations for large sets of genes (from 33.1% to 43.1% of the expressed genes in the analyzed datasets). Interestingly, we noted that the genes whose expression quantification differ depending on the reference genome used for read mapping tend to have longer exons in RefSeq annotations. Such discrepancies were also found in the sets of differentially expressed genes, subsequently impacting their biological interpretation in a Gene Ontology term enrichment analysis. With the spatial RNA-seq dataset, the choice of reference genome also had a significant impact on the generation of count matrices regarding both the number of expressed genes and their counts.

Overall, these results make the RefSeq annotations of the new assembly the most complete reference genome of *Rattus norvegicus* species for the analysis of RNA-seq data from hippocampus tissues. The analysis of spatial RNA-seq data with low sequencing depth can be impacted by the choice of reference genome at the very beginning of the analysis workflow during the count matrix generation.

References

[1] V. Marx. Method of the Year: spatially resolved transcriptomics. *Nat Methods*, 18(1):9–14, 01 2021.

[2] K. Howe, M. Dwinell, M. Shimoyama, C. Corton, E. Betteridge, A. Dove, M. A. Quail, M. Smith, L. Saba, R. W. Williams, H. Chen, A. E. Kwitek, S. A. McCarthy, M. Uliano-Silva, W. Chow, A. Tracey, J. Torrance, Y. Sims, R. Challis, J. Threlfall, and M. Blaxter. The genome sequence of the Norway rat, Rattus norvegicus Berkenhout 1769. *Wellcome Open Res*, 6:118, 2021.

[3] H. O'Leary, L. Vanderlinden, L. Southard, A. Castano, L. M. Saba, and T. A. Benke. Transcriptome analysis of rat dorsal hippocampal CA1 after an early life seizure induced by kainic acid. *Epilepsy Res*, 161:106283, 03 2020.

[4] G. Smith, A. Rani, A. Kumar, J. Barter, and T. C. Foster. Hippocampal Subregion Transcriptomic Profiles Reflect Strategy Selection during Cognitive Aging. *J Neurosci*, 40(25):4888–4899, 06 2020.

[5] C. Chen, H. Le, and C. T. Goudar. Evaluation of two public genome references for chinese hamster ovary cells in the context of rna-seq based gene expression analysis. *Biotechnol Bioeng*, 114(7):1603–1613, 07 2017.

TnSeek : analysing Tn-seq data for multiple conditions and multiple species

Loïc COUDERC[1], Erwan GUEGUEN[2], Maxime BRUNIN[1], Areski FLISSI[3], Guillemette MAROT[4], Guy CONDEMINE[2], Hélène TOUZET[3]

1 Univ. Lille, CNRS, Inserm, CHU Lille, Institut Pasteur de Lille, US 41 - UAR 2014 - PLBS - Plateforme bilille, F-59000 Lille, France
2 Univ. Lyon, Univ. Lyon 1, UMR CNRS 5240, Villeurbanne, France
3 Univ. Lille, CNRS, Centrale Lille, UMR 9189 CRIStAL, F-59000 Lille, France
4 Univ. Lille, CHU Lille, ULR 2694 - METRICS - Inria MODAL , F-59000 Lille, France

Corresponding Author: helene.touzet@univ-lille.fr

Transposon sequencing (Tn-seq) is a genome-wide screening method that identifies essential genes related to a specific phenotype or growth condition in bacteria. It consists in random transposon mutagenesis in tens of thousand cells in parallel followed by a phenotypic screen and high-throughput sequencing [1,2].

The bioinformatics analysis then typically involves several steps to identify essential genes: processing the raw sequence reads, mapping them on a reference genome, computing the fitness of each gene (including bias correction and normalisation). Several computational methods have already been proposed to perform those tasks, such as TRANSIT, ESSENTIAL, MAGenTA or Tn-Seq EXPLORER. Those tools can successfully compare datasets between two conditions and from DNA samples prepared from the same mutant pool (i.e from the same genome). However, none of them is able to deal with large projects, involving several organisms for multiple growth conditions.

In this context, we have developed a comprehensive pipeline, called TnSeek, that covers all the tasks mentioned above for one strain in one condition, and also supports comparisons of multiple conditions and identification of cross-species conserved essential genes. Another distinctive feature of TnSeek is that it includes several visualisation modules that allow to follow the analysis step by step.

This pipeline is freely-available at https://gitlab.cristal.univ-lille.fr/bonsai/tnseek as a Docker image. It is written in Python and R. It offers export files in BAM, BAI, CSV, GFF. To process the execution of the pipeline (from FASTq files to predictions) and the exploration of the results, we developed two interactive web applications. Reports in R Markdown and HTML files are generated. These two web interfaces are implemented using the Shiny R package and the SQLite database.

TnSeek is currently used in the TnPhyto project, whose objective is to identify the pathogenic core genome and strain specific genes involved in the pathogenicity of three bacterial species causing plant disease in France, *Pseudomonas syringae* and the pectinolytic enterobacteria, *Dickeya* and *Pectobacterium*.

Funding

This work is supported by ANR (TnPhyto, ANR-19-CE35-0016)

References

[1] Kwon, Y.M., Ricke, S.C. & Mandal, R.K. Transposon sequencing: methods and expanding applications. *Appl Microbiol Biotechnol* 100, 31–43 (2016). https://doi.org/10.1007/s00253-015-7037-8

[2] Cain AK, Barquist L, Goodman AL, Paulsen IT, Parkhill J, van Opijnen T. A decade of advances in transposon-insertion sequencing. *Nat Rev Genet*. 2020;21(9):526-540. https://doi:10.1038/s41576-020-0244-x

PDXploR, a transcriptomic comparison methodology for PDX models and matched Patient samples, a case study on osteosarcoma.

Robin Droit1*, Maria Eugenia Marques da Costa1*, Anne Gomez-Brouchet2, Jean-Yves Scoazec3, Audrey Mohr1, Tiphaine Adam-de-Beaumais1, Marlène Pasquet4, Birgit Geoerger1,5, Antonin Marchais1**, Nathalie Gaspar1,5**

1 INSERM U1015, Gustave Roussy, Villejuif, F-94805 France

2 Department of Pathology, IUCT-Oncopole, CHU of Toulouse and University of Toulouse; Pharmacology and structural biology institute, CNRS UMR5089, 31059 Toulouse, France

3 Department of medical Biology and pathology, Gustave Roussy, 94 805 Villejuif, France

4 Department of pediatric hemato-oncology, CHU of Toulouse, 31059 Toulouse, France

5 Department of Oncology for Children and Adolescents, Gustave Roussy, 94805 Villejuif, France

* Both author have contributed equally to the first author position

**Both author have contributed equally to the last author position

Corresponding Author: rdroit.pro@gmail.com

Osteosarcoma is the most common bone cancer in adolescents and young adults[1]. It is characterized by the formation of tumors on the metaphysis of long bones and by highly heterogenous genomic and transcriptomic profiles[2], which is complicating exploratory research on this pathology. Also, osteosarcoma is known to have a microenvironment conserved through relapse[3]. Recent research on the stratification of the osteosarcoma based on the composition of this microenvironment[4] comfort the importance of the microenvironment in this pathology.

Osteosarcoma is a rare disease[5] with a limited number of available patient samples. Therefore, bioinformatics analysis, such as classic differential expression analysis, are often inconsistent, with a low signal to noise ratio and prone to the dimensional curse.

To remediate to the dimension problematic and to study the microenvironment of this pathology, one solution is to use Patient Derived Xenograft. In the last years, patient-derived xenograft (PDX) models have been developed to better mimic the biology and heterogeneity of human tumors. PDX models have been shown to closely recapitulate the genetic alterations present in the tumor of origin but their transcriptomic landscape has been either rarely explored or analyzed with non-standardized methodology. Tumor grafting in mice constitutes a new microenvironment to colonize which could, alike metastatic dissemination, alter drastically the transcriptomic program of tumors and preclude to their preclinical significance.

Here, we propose, PDXploR a standardized method to explore the tumoral and microenvironmental transcriptomic landscapes of PDX samples. We present our result obtained for a deep cohort of 8 Osteosarcoma patients from diagnosis to relapse and their PDX avatars.

References

[1] Misaghi A, Goldin A, Awad M, Kulidjian AA. Osteosarcoma: a comprehensive review. *SICOT J*. 2018;4:12. doi:10.1051/sicotj/2017028

[2] Rickel K, Fang F, Tao J. Molecular genetics of osteosarcoma. *Bone*. 2017;102:69-79. doi:10.1016/j.bone.2016.10.017

[3] Zhou, Y., Yang, D., Yang, Q. *et al.* Single-cell RNA landscape of intratumoral heterogeneity and immunosuppressive microenvironment in advanced osteosarcoma. *Nat Commun* **11,** 6322 (2020). https://doi.org/10.1038/s41467-020-20059-6

[4] Antonin Marchais, Maria Eugenia Marques Da Costa, Nathalie Gaspar et al.; Immune infiltrate and tumor microenvironment transcriptional programs stratify pediatric osteosarcoma into prognostic groups at diagnosis. *Cancer Res* 2022; https://doi.org/10.1158/0008-5472.CAN-20-4189

[5] Ward E, DeSantis C, Robbins A, Kohler B, Jemal A. Childhood and adolescent cancer statistics, 2014. CA Cancer J Clin. 2014 Mar-Apr;64(2):83-103. doi: 10.3322/caac.21219. Epub 2014 Jan 31. PMID: 24488779.

Prediction of pediatric cancer patient progression with non-invasive procedures: Pipeline to automatize liquid biopsy analysis

B. Audinot[1,2], A. Mohr[1,2], K. Massau[1,2], M. Jimenez[3], N. Gaspar[1,2], A. Marchais[1,2], S. Abbou[2]

1 BiiOSTeam, U1015, Gustave Roussy, 94805, Villejuif, France
2 Department of Oncology for Children and Adolescents, Gustave Roussy, 94805 Villejuif, France
3 R&D Unicancer, Paris, France

Liquid biopsy in oncology emerged over the last years as an efficient non-invasive procedure to evaluate disease progression. Through liquid biopsy, clinicians get access to circulating molecules originating from the tumor and, likewise, infer the tumor state over therapeutic interventions. In Sarcoma, quantification of blood circulating tumor DNA (ctDNA) is performed by the estimation of specific (Somatic mutation, Translocation) or global structural alterations (Copy Number Alterations) of the tumor genome.

There, we developed and evaluated a pipeline to automatize the analysis of blood liquid biopsies from 170 Osteosarcoma patients through time. Benchmark of available methodologies to estimate ctDNA fractions from low coverage Whole Genome Analysis (lcWGS) allowed us to optimize and automatize their analysis and anticipate the deployment in a clinical setup. In this study, we demonstrate that ctDNA fraction estimated at diagnosis in Osteosarcoma is a prognosis factor to predict the disease progression that improve the current multivariate model based until now on clinical features (Metastasis at diagnosis, Treatment response). Finally, we propose to the community a shiny interface to predict the progression risk with their own ctDNA fraction data.

References

1. Adalsteinsson, V. A. *et al.* Scalable whole-exome sequencing of cell-free DNA reveals high concordance with metastatic tumors. *Nat Commun* **8**, 1324 (2017).
2. Raman, L., Dheedene, A., De Smet, M., Van Dorpe, J. & Menten, B. WisecondorX: improved copy number detection for routine shallow whole-genome sequencing. *Nucleic Acids Research* **47**, 1605–1614 (2019).
3. Misaghi A, Goldin A, Awad M, Kulidjian AA. Osteosarcoma: a comprehensive review. *SICOT J*. 2018;4:12. doi:10.1051/sicotj/2017028
4. Rickel K, Fang F, Tao J. Molecular genetics of osteosarcoma. *Bone*. 2017;102:69-79. doi:10.1016/j.bone.2016.10.017

Single-cell deconvolution model predictive of patient survival in clear cell Renal Cell Carcinoma (ccRCC)

Gwendoline LECUYER[1], Judikaël SAOUT[1], Bertrand EVRARD[1], Paul RIVAUD[1], Simon LÉONARD[1,2,3], Nathalie RIOUX-LECLERCQ[1], Aurélie LARDENOIS[1] and Frédéric CHALMEL[1]

[1]Univ Rennes, Inserm, EHESP, Irset (Institut de recherche en santé, environnement et travail) – URM_S 1085, F-35000 Rennes, France
[2]UMR 1236, University of Rennes, INSERM, Etablissement Français du Sang Bretagne, Rennes, France
[3]LabEx IGO "Immunotherapy, Graft, Oncology", Nantes, France

Corresponding Author: gwendoline.lecuyer@univ-rennes1.fr

Inter- and intra-tumor heterogeneity (ITH) are known to be a crucial determinant of therapeutic resistance and treatment failure, and one of the main reasons for poor overall survival in metastatic patients.[1] Over the last few years, ITH has been extensively described in many types of tumors with single-cell transcriptomics approaches. The latter make it possible to characterize all the tumoral clones but also to study their evolution in order to understand and predict tumor progression, metastases and therapeutic responses.[2] ccRCC patients with high risk of recurrence receive an adjuvant therapy to limit the risk of development of metastases. Nevertheless, among the patients classified as "low-risk", 7% will relapse during follow-up.[3,4] In this context, we developed a project aiming at studying ccRCC tumor cell evolution by integrating published single-cell RNA-seq (scRNA-seq) data (Young et *al*, 2018) with data generated within our laboratory. After lineage analysis, we were able to identify a patient with several populations of cancer cells that mark the different stages of epithelial-mesenchymal transition (EMT) and of tumorigenesis. These cell populations were then predicted in bulk RNA-seq data from 525 ccRCC patients (TCGA database https://www.cancer.gov/tcga), using a deconvolution approach. Our first results seem to indicate a very significant association between the proportions of some identified cell populations with the survival of patients. Ultimately, the objective is to improve the prediction of the patient's risk of relapse and to refine "low-risk" patient stratification based on cancer cell population proportion in the tumor with the long-term goal of adapting patient care in a personalized way.

References

[1] S. M. Crusz *et al.*, « Heterogeneous response and progression patterns reveal phenotypic heterogeneity of tyrosine kinase inhibitor response in metastatic renal cell carcinoma », *BMC Med.*, vol. 14, nº 1, p. 185, déc. 2016, doi: 10.1186/s12916-016-0729-9.

[2] C. Krishna *et al.*, « Single-cell sequencing links multiregional immune landscapes and tissue-resident T cells in ccRCC to tumor topology and therapy efficacy », *Cancer Cell*, vol. 39, nº 5, p. 662-677.e6, mai 2021, doi: 10.1016/j.ccell.2021.03.007.

[3] B. Rini *et al.*, « A 16-gene assay to predict recurrence after surgery in localised renal cell carcinoma: development and validation studies », *Lancet Oncol.*, vol. 16, nº 6, p. 676-685, juin 2015, doi: 10.1016/S1470-2045(15)70167-1.

[4] B. Kang *et al.*, « T1 Stage Clear Cell Renal Cell Carcinoma: A CT-Based Radiomics Nomogram to Estimate the Risk of Recurrence and Metastasis », *Front. Oncol.*, vol. 10, p. 579619, nov. 2020, doi: 10.3389/fonc.2020.579619.

Multi-omics and multi-tissues data to improve the understanding of heat stress adaptation mechanisms

Guilhem Huau[1,2], David Renaudeau[1], Jean-Luc Gourdine[3], Katia Fève[2], Yannick Lippi[4], Juliette Riquet[2] and Laurence Liaubet[2]

[1] PEGASE, INRAE, Institut Agro, 35590, Saint Gilles, France
[2] GenPhySE, Université de Toulouse, INRAE, ENVT, 31326, Castanet Tolosan, France
[3] URZ, INRAE, Domaine Duclos Prise d'eau, Petit-Bourg, France
[4] INRA UMR1331, ToxAlim, University of Toulouse, Toulouse, France

Corresponding author: guilhem.huau@inrae.fr

Heat stress is one of the main limiting performance factors in the pig industry, and its importance is expected to grow with the consequences of warming climate. Different solutions to temper the influence of heat stress on the performance and well-being of pigs have been studied [1], and the genetic selection appear as a potential solution. Genetic mechanisms behind heat tolerance are still not well understood, but some differences between breeds have been shown [2], as well as an interaction between genetic and environment [3].

This study aims to use multi-omics and multi-tissues data to understand the genetic pathways involved in heat stress adaptation. To achieve this, transcriptomic data from seven tissues, metabolomic data from four tissues and blood parameters where obtained in an experiment involving 36 pigs from 3 breeds, slaughtered before (n=18) or after a 5-d exposure to 32°C (n=18). With the aim to identify differentially expressed genes (DEG) and differentially produced metabolites, data were analysed with a mixed model with the effects of HS conditions, breed and sire origin. Then using a sPLSda multi-block (with the framework DIABLO of the mixOmics R package [4]) as an integrative method, we perform a multi-tissues and multi-omics analysis to identify the interactions involved in heat stress adaptation at a genetic level.

We identified 12912 unique DEG in all tissues. From the enrichment analysis using the R package gProfiler [5], querying GO and KEGG databases, we have identified recurrent enrichment terms, such as energy metabolism and cell cycle, but also tissues specific terms. Integrating data from the different tissues, we have identified co-expression network of genes involved in heat tolerance.

In the future, we plan to complement this study with other datasets, involving longitudinal data and larger cohort.

Acknowledgements

This work is funded by MP ACCAF projet P10157 (PigChange). The PhD project is funded by INRAE and DGER.

References

[1] D. Renaudeau, A. Collin, S. Yahav, V. de Basilio, J. L. Gourdine, and R. J. Collier. Adaptation to hot climate and strategies to alleviate heat stress in livestock production. *Animal: An International Journal of Animal Bioscience*, 6(5):707–728, May 2012.

[2] R. Rosé, H. Gilbert, T. Loyau, M. Giorgi, Y. Billon, J. Riquet, D. Renaudeau, and J.-L. Gourdine. Interactions between sire family and production environment (temperate vs. tropical) on performance and thermoregulation responses in growing pigs1,2. *Journal of Animal Science*, 95(11):4738–4751, November 2017.

[3] J.L. Gourdine, J. Riquet, R. Rosé, N. Poullet, M. Giorgi, Y. Billon, D. Renaudeau, and H. Gilbert. Genotype by environment interactions for performance and thermoregulation responses in growing pigs,. *Journal of Animal Science*, 97(9):3699–3713, September 2019.

[4] F. Rohart, B. Gautier, A. Singh, and K-A L. Cao. mixOmics: An R package for 'omics feature selection and multiple data integration. *PLOS Computational Biology*, 13(11):e1005752, November 2017.

[5] L. Kolberg, U. Raudvere, I. Kuzmin, J. Vilo, and H. Peterson. Gprofiler2 – an R package for gene list functional enrichment analysis and namespace conversion toolset g:Profiler, November 2020.

Single-cell ATAC-seq integration highlight epigenetics remodeling within HSC quiescence signaling.

Alexandre PELLETIER[1], Fabien DELAHAYE[1] and Philippe FROGUEL[1]

[1] UMR 1283 EGID – CHRU de Lille, Faculté de Médecine HENRI-WAREMBOURG, Pôle Recherche : 1, place de Verdun, 59045 LILLE CEDEX – France

Corresponding Author: alexandre.pelletier.etu@univ-lille.fr

Fetal overgrowth increases the risk to develop chronic diseases like type 2 diabetes and cardiovascular diseases later in life, but the mechanism involved remains unclear [1]. The hematopoietic system plays a critical role in processes like inflammation, immunity, and cardiovascular repair throughout life, making its progressive alteration a candidate mechanism in the development of these diseases[2]. Our previous work found a global DNA hypermethylation in hematopoietic stem and progenitors cells (HSPCs) from Large for Gestational Age neonates (LGA), supporting an early epigenetic alteration of the hematopoietic system [3].

Here, through multimodal analysis, we investigate the influence of early programming on the regulatory landscape of HSPCs. We added key single cell chromatin accessibility (scATAC-seq) data on our previous bulk DNA methylation and single cell transcriptomic analysis of LGA compared to appropriately grown neonates (CTRL) HSPCs. Using both TF-targets gene co-expression and TF motif accessibility data, we found that the EGR1-KLF2 regulatory network, known to regulate proliferation and differentiation of Hematopoietic stem cells (HSCs), was specifically affected by epigenetic and transcriptomic alterations. Indeed, this regulatory network was enriched for open chromatin regions with decreased accessibility, DNA hypermethylation, as well as for downregulated genes in LGA HSCs supporting functional consequences on quiescence regulation. Finally, we implemented a generative adversarial network (GAN) framework based on variational auto-encoder and batch discriminator neural networks [5] to improve integration of scATAC-seq and scRNA-seq data and find key cis regulatory elements associated with epigenetics alterations triggering transcriptional and functional changes.

Taken together, our study support a model where foetal overgrowth affect HSC quiescence signalling through an epigenetics programming of the EGR1-KLF2 related regulatory network, and provide an analytical tool to decipher epigenetics mechanism associated with long-term tissue dysfunctions and disease susceptibility.

References

1. Evagelidou E, Giapros I, Challa A, et al. Prothrombotic State, Cardiovascular, and Metabolic Syndrome Risk Factors in Prepubertal Children Born Large for Gestational Age. *Diabetes Care*. 2010
2. Pardalia E, Dimmeler S, Zeiher A, Rieger M. Clonal hematopoiesis, aging, and cardiovascular diseases. *Experimental Hematology*. 2020
3. Delahaye F, Wijetunga NA, Heo HJ, et al. Sexual dimorphism in epigenomic responses of stem cells to extreme fetal growth. *Nat Commun*. 2014
4. Bahrami M, Maitra M, Nagy C, et al. Deep feature extraction of single-cell transcriptomes by generative adversarial network. *Bioinformatics*. 2021

Investigating the resistance to mIDH inhibitors in Acute Myeloid Leukemia integrating multiple regulatory layers

Alexis Hucteau[1], Nathaniel Polley[1], Jean-Emmanuel Sarry[1] and Vera Pancaldi[1]

[1] INSERM, Centre de Recherches en Cancérologie de Toulouse (CRCT, Toulouse, France)
Corresponding Author: alexis.hucteau@inserm.fr

The mutation in the gene isocitrate dehydrogenase 1 (IDH1) is implicated in Acute Myeloid Leukemia (AML), as cells with the alteration abnormally produce an oncometabolite 2-hydroxyglutarate (2-HG). 2-HG was found to cause widespread changes in DNA methylation [1]. IDH inhibitors have shown good clinical response in AML patients. However, primary or acquired resistance to IDH inhibitor therapies represent a major problem limiting their efficacy. The mechanisms that mediate resistance to IDH inhibition are poorly understood.

To have a complete idea of the dysregulations that happen in AML cells, we need to take into account the widespread effects of the mutation. The blockade of the differentiation that characterizes AML is associated with epigenomic alterations that can change both DNA methylation profiles (as is known in IDH mutants) and chromatin's 3D organization, involving contacts between genes and their regulatory elements. Methylation on both enhancers and promoters can repress or activate gene expression by blocking or enhancing the recruitment of transcription factors. Here, we present an analysis of gene expression and DNA methylation taking into account chromatin structure networks to explore specific gene regulation mechanisms involved in resistance to IDH inhibitors.

Exploiting published 3D blood cells epigenomes, we are able to associate each group of AML patients harboring an IDH mutation from our cohort [4] with the chromatin structure network of the closest blood cell type. We then use the reference chromatin contact network assigned to either responders, non-responders and relapse samples and investigate the 3D epigenome relation between gene expression and methylation alterations.

Furthermore, we infer gene regulatory networks from gene expression data [5] and predict TF activities [6] to reconstruct a specific transcriptional regulatory network associated with resistance. Finally, we combine this regulatory network with a protein - protein interactions to consider further regulatory levels. To better understand the connection between transcriptional regulation and metabolic alterations, we will also add a layer related to metabolic pathways activity estimated by combining gene expression data with a full genome-scale model of metabolism [7].

These different regulatory layers will be integrated into a multiplex network (a multilayer structure where some of the nodes are in common between layers and connections between nodes in each layer exist [8]) to find key elements that are involved in the resistance, to ultimately propose new targets and therapeutic approaches to improve efficacy in patients.

References

1. Maria E Figueroa & al. DNA methylation signatures identify biologically distinct subtypes in acute myeloid leukemia, Cancer Cell, 2010.
2. Javierre BM & al, Lineage-Specific Genome Architecture Links Enhancers and Non-coding Disease Variants to Target Gene Promoters, Cell, 2016.
3. Miguel Madrid-Mencía, Vera Pancaldi & al, Using GARDEN-NET and ChAseR to explore human haematopoietic 3D chromatin interaction networks, Nucleic Acids Research, 2020
4. Feng Wang, Courtney Dinardo, Koichi Takahashi & al, Leukemia stemness and co-occurring mutations drive resistance to IDH inhibitors in acute myeloid leukemia, Nature Communication, 2021.
5. Margolin, A.A., Nemenman, I., Basso, K. et al. ARACNE: An Algorithm for the Reconstruction of Gene Regulatory Networks in a Mammalian Cellular Context. BMC Bioinformatics, 2006.
6. Alvarez MJ, Shen Y, Giorgi FM, Lachmann A, Ding BB, Ye BH & Califano, A. Functional characterization of somatic mutations in cancer using network-based inference of protein activity. Nature Genetics, 2016.
7. J. L. Robinson, P. Kocabaş, H. Wang, P.-E. Cholley, et al. An atlas of human metabolism. Sci. Signal. 2020.
8. Cozzo, Emanuele & Ferraz de Arruda, Guilherme & Rodrigues, Francisco & Moreno, Yamir. Multiplex Networks: Basic Definition and Formalism, 2018.

Detection of nucleotide repeat expansions by exome sequencing of Parkinson's disease patients using ExpansionHunter

FANNY CASSE [1], THOMAS COURTIN [1,3], CHRISTELLE TESSON [1], MÉLANIE FERRIEN [1], SANDRINE NOEL [3], ANNE-LAURE FAURET AMSELLEM [3], THOMAS GAREAU [2], JUSTINE GUEGAN[2], SUZANNE LESAGE [1], JEAN-CHRISTOPHE CORVOL [1,4], AND ALEXIS BRICE [1]

[1] Corti-Corvol Team, Institut du Cerveau-Paris Brain Institute-ICM,Sorbonne Université, Hôpital de la Pitié-Salpêtrière, Inserm U 1127, CNRS UMR 7225, 47 boulevard de l'Hôpital,75013, Paris, France.

[2] Data Analysis Core, Institut du Cerveau-Paris Brain Institute-ICM,Sorbonne Université, Hôpital de la Pitié-Salpêtrière, Inserm U 1127, CNRS UMR 7225, 47 boulevard de l'Hôpital,75013, Paris, France.

[3] Département de Génétique, APHP, Hôpital de la Pitié Salpêtrière, 47 boulevard de l'Hôpital, 75013, Paris, France.

[4] Département de Neurologie, APHP, Hôpital de la Pitié Salpêtrière, 47 boulevard de l'Hôpital, 75013, Paris, France

Corresponding Author: fanny.casse@icm-institute.org

Short Tandem Repeats represent 10% of the human genome. Beyond a threshold specific to each locus, a progressive disease manifest which very often affects the central nervous system. Repeat expansion in the *ATXN2* gene is known to be responsible for spinocerebellar ataxia type 2, but when the expansion is interrupted it can cause Parkinson's Disease (PD) [1,2]. Repeat sequence expansions in PD patients have also been found in the *ATXN3* gene. High-throughput whole exome sequencing (WES) plays an important role in the identification of monogenic forms of PD[3] but detects mostly point mutations or small insertions/deletions.

In this context, we analyzed WES data from our cohort of PD patients enriched in familial and early-onset forms (under 45 years of age), previously excluded from major PD genes, using ExpansionHunter [4,5] software. This software allows to estimate the size of the repeats in the target genes from alignment files.

The putative expansions were then validated by genotyping. Our cohort consisted of 827 WES of PD patients performed with Rochev3 (n=213), MedExome (n=171) or Twist (n=443) enrichment kits. The mean size estimates of the expansions for *ATXN2* (22 CAG) and *ATXN3* (19 CAG) are consistent with the Caucasian population means repeat size[6,7]. Our bioinformatics analysis allowed us to identify 1 patient who had a repeat size within the pathological confidence range (>55 CAG repeats) in *ATXN3* and 3 patients with repeat sizes within the pathological range (>32 CAG repeats) in *ATXN2*, 2 of whom belonged to the same family. No pathogenic repeats were found on other genes with a sufficient coverage for analysis. Sequencing analysis of the *ATXN2* repeat in 3 patients has revealed the presence of 4 CAA interruptions.

These results demonstrate the usefulness of bioinformatic detection of expansions using ExpansionHunter in exome sequencing data of PD patients. They also underline the fact that *ATXN2* expansions with more than three interruptions are not a rare cause of PD, particularly among dominant families (1% in our cohort).

References

1. Charles, P. *et al.* Are interrupted SCA2 CAG repeat expansions responsible for parkinsonism? *Neurology* **69**, 1970–1975 (2007).
2. Lu, C.-S., Wu Chou, Y.-H., Kuo, P.-C., Chang, H.-C. & Weng, Y.-H. The parkinsonian phenotype of spinocerebellar ataxia type 2. *Arch. Neurol.* **61**, 35–38 (2004).
3. Bras, J. M. & Singleton, A. B. Exome sequencing in Parkinson's disease. *Clin. Genet.* **80**, 104–109 (2011).
4. Dolzhenko, E. *et al.* ExpansionHunter: a sequence-graph-based tool to analyze variation in short tandem repeat regions. *Bioinforma. Oxf. Engl.* **35**, 4754–4756 (2019).
5. Dolzhenko, E. *et al.* Detection of long repeat expansions from PCR-free whole-genome sequence data. *Genome Res.* **27**, 1895–1903 (2017).
6. Imbert, G. *et al.* Cloning of the gene for spinocerebellar ataxia 2 reveals a locus with high sensitivity to expanded CAG/glutamine repeats. *Nat. Genet.* **14**, 285–291 (1996).
7. Sobczak, K. & Krzyzosiak, W. J. CAG repeats containing CAA interruptions form branched hairpin structures in spinocerebellar ataxia type 2 transcripts. *J. Biol. Chem.* **280**, 3898–3910 (2005).

Integration analysis for AAV-based gene therapy vectors with linked-read sequencing

Mallaury VIE[1,2], Louisa JAUZE[1,2], Tiziana LA BELLA[1,2], Kevin CHEESEMAN[3], Julien COTTINEAU[3] and Giuseppe RONZITTI[1,2]

[1] Genethon, 91000, Evry, France

[2] Université Paris-Saclay, Univ Evry, Inserm, Integrare research Unit UMR_S951, 91000, Evry, France

[3] Whitelab Genomics, 91000, Evry, France

Corresponding Authors: mvie@genethon.fr, gronzitti@genethon.fr, kcheeseman@whitelabgx.com

Gene therapy is a technique that uses a gene to treat, prevent or cure a disease. The mechanism involves replacing a disease-causing gene with a healthy copy of the gene. Due to their natural ability to infect cells, vectors of viral origin are one of the most efficient methods to deliver genetic payloads into cells. In order to make them safe to use, most of the viral genes are removed, and the vector is modified to deliver only therapeutic genes. Recombinant adeno-associated viral (rAAV) vectors are considered promising tools for gene therapy and are already used in clinical trials and approved drugs. While rAAV is thought to be an episomal vector, observations have shown that it can integrate in the cell host genome at low frequency [1]. rAAV integrations occur randomly and contain partial elements, deletions and rearrangements of the vector [2]. Rodent studies have shown that integration can occur in tumors, suggesting some link between rAAV integration and genotoxic events [3]. For now, there is no evidences of genotoxicity in humans, but a study reported the association between wild-type virus integration and insertional mutagenesis of cancer-driver genes in tumors [4].

To perform integration site analysis, popular methods based on short-read sequencing like LAM-PCR [5] or target enrichment sequencing [6] are available. They rely on the detection of chimeric read (vector-host) to identify an integration site, but they do not give any information on the structure of the vector integrants. Other methods based on long-read sequencing like PacBio or Nanopore can overcome this limitation but lose accuracy and are expensive. New methods, based on linked-read sequencing like TELL-Seq [7], may provide the long range information missing in short-read sequencing. TELL-Seq relies on the use of a unique barcode to tag every short-read generated from the same long DNA molecule. This allows to deduce that those reads came from the same molecule.

This work aims at better characterize integration events with the help of the linked-read sequencing to overcome the biases of the standard methods. The samples used are coming from mice that received AAV vector-based gene therapy treatment for the correction of a metabolic liver disorder with underlying tumor development. We used TELL-Seq method to compare the integration profile of rAAV in tumor and non-tumor samples in order to determine the vector elements of the viral genome involved in these integration events. We identified some possible integration sites but the validation seems complicated. Nevertheless, this technique has some potential and further work is still ongoing to overcome the existing limitations.

Acknowledgements

This PhD is funded by a grant from the DIM Gene Therapy (Paris Region PhD 2020).

References

[1] Donsante A et al. Observed incidence of tumorigenesis in long-term rodent studies of rAAV vectors. *Gene Ther*. 8(17):1343-6, 2001.

[2] Nguyen GN et al. A long-term study of AAV gene therapy in dogs with hemophilia A identifies clonal expansions of transduced liver cells. *Nat Biotechnol*. 39(1):47-55, 2021.

[3] Chandler RJ et al. Vector design influences hepatic genotoxicity after adeno-associated virus gene therapy. *J Clin Invest*. 125(2):870-80, 2015.

[4] Nault JC et al. Recurrent AAV2-related insertional mutagenesis in human hepatocellular carcinomas. *Nat Genet*. 47(10):1187-93, 2015.

[5] Schmidt M et al. High-resolution insertion-site analysis by linear amplification-mediated PCR (LAM-PCR). *Nat Methods*. 4(12):1051-7, 2007.

[6] Ohnuki N et al. A target enrichment high throughput sequencing system for characterization of BLV whole genome sequence, integration sites, clonality and host SNP. *Sci Rep*. 11(1):4521. 2021.

[7] Chen Z et al. Ultralow-input single-tube linked-read library method enables short-read second-generation sequencing systems to routinely generate highly accurate and economical long-range sequencing information. *Genome Res*. 30(6):898-909, 2020.

Interaction dynamics comparison of the models of Omicron BA.1 and BA.2 variants of SARS-CoV-2 Spike RBD in complex with human ACE2 through MD simulations and MM-PBSA calculations

Audrey DEYAWE KONGMENECK[1], Mariem GHOULA[1], Gautier MOROY[1] and Anne-Claude CAMPROUX[1]

[1] Université de Paris, BFA, UMR 8251, CNRS, ERL U1133, Inserm, F-75013 Paris, France

Corresponding Author: anne-claude.camproux@u-paris.fr

The severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) is the pathogen that gave rise to the COVID-19 outbreak [1]. Originally discovered in Wuhan, China, in late December 2019, the epidemic was quickly labeled a pandemic by the World Health Organization (WHO) in March 2020. The SARS-CoV-2 has turned out to be an extremely unstable virus [2]. Thus, since December 2020, four new variants of the SARS-CoV-2 virus have emerged in the UK, South-Africa, Brazil, and India. Each variant was classified one after another as a Variant of Concern (VOC) by the WHO. In late November 2021, a fifth variant, called Omicron, arose and then was classified along with its five sub-lineages as a VOC by the WHO. To date, the most infectious sub-lineages of the Omicron variant are BA.1 and BA.2 [3]. As by SARS-CoV virus, the Angiotensin Converting Enzyme II (ACE2) is used as a host-cell receptor by SARS-CoV-2 virus [4]. Indeed, the entry of coronaviruses is usually mediated by their Spike (S) protein that forms homotrimers. Each S monomer carries a conserved Receptor Binding Domain (RBD) that interacts directly with ACE2 [5]. Despite the design of vaccines and monoclonal antibodies to prevent SARS-CoV-2 Spike from interacting with ACE2, no treatment has been found to treat the COVID-19 disease regardless of the causing variant [6]. To that extent, several structural studies, including experimental and computational methods, have been used in order to shed light into the binding interface of SARS-CoV-2 Spike RBD and ACE2 proteins [7]. Specifically, Molecular Mechanics Generalized-Born Surface Area free-energy calculations, conducted on Molecular Dynamics (MD) trajectories of models of SARS-CoV and SARS-CoV-2 Spike RBD, each in complex with ACE2, allowed to study their conformational space, as well as to predict the free-energies associated with ACE2 binding. Results showed that SARS-CoV-2 Spike RBD binds ACE2 with a higher energy than SARS-CoV. Similarly, our recent computational studies showed that, in comparison to the Wild-Type (WT) strain of SARS-CoV-2, the VOC turned out to bind ACE2 with free energies that increase with their respective date of appearance. To that extent, we used MD simulations and Molecular Mechanics Poisson Boltzmann Surface Area (MM-PBSA) free-energy calculations to investigate the structure [8] of the complexes formed by ACE2 and the RBD of BA.1 and BA.2 Omicron sub-lineages. Altogether, these computational approaches allowed to determine the molecular determinants that underlie the higher binding energies of each BA.1 and BA.2 variants of SARS-CoV-2 Spike RBD protein with ACE2 in comparison with the WT.

References

1. Chen N, et al., Epidemiological and clinical characteristics of 99 cases of 2019 novel coronavirus pneumonia in Wuhan, China: a descriptive study. *Lancet*, 395:507–513, 2020
2. Giovanetti M, et al., Evolution patterns of SARS-CoV-2: Snapshot on its genome variants. *Biochem Biophys Res Commun*, 538:88–91, 2021
3. Chen J and Wei G-W, Omicron BA.2 (B.1.1.529.2): high potential to becoming the next dominating variant. *Res Sq* 13:3840–3849, 2022
4. Zhou P, et al., (2020) A pneumonia outbreak associated with a new coronavirus of probable bat origin. *Nature* 579:270–273, 2020
5. Lan J, et al., (2020) Structure of the SARS-CoV-2 spike receptor-binding domain bound to the ACE2 receptor. *Nature* 581:215–220, 2020
6. Piccoli L et al., Mapping Neutralizing and Immunodominant Sites on the SARS-CoV-2 Spike Receptor-Binding Domain by Structure-Guided High-Resolution Serology. *Cell* 183:1024-1042, 2020
7. Spinello A, et al., Is the Rigidity of SARS-CoV-2 Spike Receptor-Binding Motif the Hallmark for Its Enhanced Infectivity? Insights from All-Atom Simulations. *J Phys Chem Lett* 11:4785–4790, 2020
8. Mannar D, Saville JW, Zhu X, et al., SARS-CoV-2 Omicron variant: Antibody evasion and cryo-EM structure of spike protein–ACE2 complex. *Science (80-)* 375:760–764, 2022

Atlas and biological significance of transcribed non-coding regions of the human genome

Pierre De Langen[1], Fayrouz Hammal[1], Lionel Spinelli[1,2] and Benoit Ballester[1]

[1] TAGC Inserm U1090, 163 avenue de Luminy, 13288, Marseille 09, France
2 CIML, 163 avenue de Luminy, 13288, Marseille 09, France

Corresponding Authors: pierre.de-langen@inserm.fr, benoit.ballester@inserm.fr

1. Introduction

A large fraction of the non-coding human genome has been shown to be transcribed [1]. These transcripts -Enhancer RNAs, ncRNAs or transposons to name a few- have been shown to play a key role in cell identity as well as in several diseases such as cancers. However, despite this biological and clinical relevance, the global identification and functional characterization of these RNAs is quite behind its coding counterpart. Most of the sequencing techniques employed to detect non-coding transcripts directly target the non-coding RNAs, which are usually less stable and less abundant than coding RNAs. Instead, we use an indirect approach, where we target the enzyme responsible for the transcription, the RNA Polymerase II (Pol II), using Chromatin Immuno-Precipitation Sequencing (ChIP-seq).

2. Results

We annotated and re-analyzed uniformly all publicly available Pol II ChIP-seq experiments (906 after quality control), from ENCODE [2] and GEO in order to establish a catalogue of Pol II bound regions (named Pol II consensus) of the genome. More precisely, we create a binary 2D matrix where rows correspond to the 906 samples and columns to the Pol II consensuses. For this study, we only focus on intergenic Pol II consensuses (excluding promoters, exons and introns).

Using an unsupervised graph clustering approach, we grouped Pol II consensuses based on their binding patterns across experiments. We identified groups of Pol II bound regions which appear to be highly specific to certain cell/tissues types. The biological relevance of these clusters is confirmed by strong enrichments in biological and clinical features for each cluster, such as enrichments in causal SNPs for tissue-related traits.

Next, we observe that in RNA-seq experiments, a majority of the intergenic signal is located at the previously identified Pol II consensuses, which suggests that our Pol II probes are good sampling points for studying the intergenic transcription. Using only the RNA-seq signal located at the intergenic Pol II probes, we are able to discriminate between different biological conditions in three different large scale RNA-seq datasets originating from TCGA [3], GTex [4] and ENCODE [2] . In the TCGA dataset, we identified intergenic markers associated either with survival or the tumour state of the tissue, at a per-cancer scale but also at a pan-cancer level.

Acknowledgements

Centre de Calcul Intensif d'Aix-Marseille is acknowledged for granting access to its high performance computing resources. PhD Fellowship to P.D.L. from the French Ministry of Higher Education and Research.

References

[1] R. Andersson *et al.*, "An atlas of active enhancers across human cell types and tissues," *Nature*, vol. 507, no. 7493, Art. no. 7493, Mar. 2014, doi: 10.1038/nature12787.
[2] C. A. Davis *et al.*, "The Encyclopedia of DNA elements (ENCODE): data portal update," *Nucleic Acids Res.*, vol. 46, no. D1, pp. D794–D801, Jan. 2018, doi: 10.1093/nar/gkx1081.
[3] J. N. Weinstein *et al.*, "The Cancer Genome Atlas Pan-Cancer analysis project," *Nat. Genet.*, vol. 45, no. 10, Art. no. 10, Oct. 2013, doi: 10.1038/ng.2764.
[4] J. Lonsdale *et al.*, "The Genotype-Tissue Expression (GTEx) project," *Nat. Genet.*, vol. 45, no. 6, Art. no. 6, Jun. 2013, doi: 10.1038/ng.2653.

Improving clinical diagnosis using Nanopore Adaptive Sampling and NanoCliD

Eleonore FROUIN[1], Kevin MERCHADOU[1], Mathilde FILSER[2], Abderaouf HAMZA[2], Elodie GIRARD[3], Nicolas SERVANT[3], Julien MASLIAH-PLANCHON[2] and Victor RENAULT[1]

1 Bioinformatique Clinique - PMDT, Institut Curie, 26 rue d'Ulm, 75005 Paris, France
2 Unité de Génétique Somatique, Institut Curie, 26 rue d'Ulm, 75005 Paris, France
3 Bioinformatics core facility, Institut Curie, INSERM U900, Mines Paris tech, PSL University, 75005 Paris, France

Corresponding Author: victor.renault@curie.fr

Keywords Adaptive sampling, long reads, diagnosis, variant phasing, SV detection, methylation

Git : https://github.com/InstituteCurieClinicalBioinformatics/NanoCliD

1. Introduction

Nanopore sequencing (Oxford Nanopore Technologies) produces long-read data that have great potential for diagnoses of cancer [1,2,3]. Indeed, the library preparation is simple and fast (1h30) and long reads can be used to identify a wide range of genetic alterations (structural variants, copy number variation, variants, variant phasing) in addition to a direct estimation of DNA methylation levels. So far, its use in the clinical diagnosis remains limited by the relative low sequencing throughput and the large amount of material required to achieve sufficient read depths [3]. These limitations can be overcome by using adaptive sampling to achieve better sequencing coverage in targeted regions without additional sample preparation such as capture or PCR amplification [4]. Nanopore adaptive sampling relies on the ability to read a DNA sequence in real time: if the read does not match the sequence of interest (deducted from the 500 first bases), the DNA molecule is ejected, leaving the pore free for another DNA molecule in the sample. Otherwise, the sequence is recognized and the sequencing continues. At Institute Curie, adaptive sampling is used as a backup to improve a diagnosis when geneticists are facing the inherent limitations of short reads sequencing. The Clinical Bioinformatics Team provides NanoCliD (Nanopore Clinical Diagnosis), a toolkit designed to capture genomic alterations including methylation profile in nanopore adaptive sampling data.

2. Material and Methods

Implemented with Snakemake, the toolkit NanoCliD processes Nanopore adaptive sequencing (targeting ~500 genes and surrounding regions). Basecalling is performed with guppy leading to FastQ files generation. Reads are then mapped to reference genome using minimap2. Four calling analyses can be performed from the BAM and FastQ files obtained from previous steps. Structural variant (SV) calling step is made of four tools listed in figure 1 for which outputs are merged with SURVIVOR. VCF is then annotated with AnnotSV and a circos plot is generated based on reported SVs. Deeptools and R package QDNASeq are used to call CNVs from BAM files. These CNVs are then summarized as a single pan-genomic CNV profile. The three tools listed in figure 1 perform variant calling. Only PEPPER and NanoCaller phase the variants. VCFs are merged using an in-house

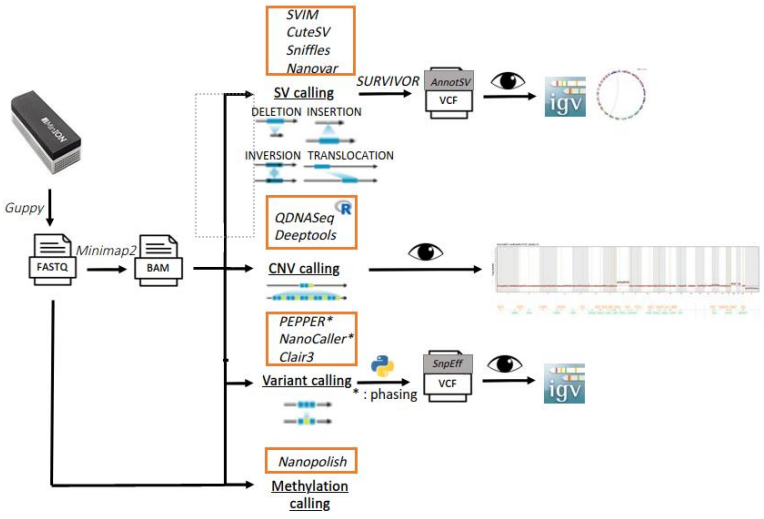


Fig 1. Workflow of NanoCliD toolkit

python script and SnpEff annotates the resulting VCF. Variants are then reviewed manually with IGV. Finally, methylation calling based on FastQ and BAM files is achieved with Nanopolish.

3. Results

NanoCliD constitutes a real help for clinical diagnosis as illustrated through the following 3 case studies. A first example shows the contribution of this multimodal toolkit for a tumor of a young patient whose diagnosis was difficult in anatomical pathology. Moreover, no characteristic gene fusion was detected in this tumor using short-read RNAseq sequencing. After adaptive sequencing, NanoCliD was able to classify this tumor as an ependymoma with a *YAP1-MAMLD1* fusion based on its methylation profile. The fusion was confirmed by the 4 structural variant detection tools. The reads supporting the fusion were validated manually on IGV. A CNV could also be detected in this tumor with a deletion of chromosome 22, a well-known alteration in ependymomas.

A second example is a patient suffering from thrombotic microangiopathy. The causing mutation is a missense variant in *CFH* gene and was reported using NGS short reads in three weeks. Using adaptive sequencing and NanoCliD, only three days were needed to report this missense variant. Besides, three other variants of interest detected in short reads were reported. Thanks to PEPPER and NanoCaller, the variants were phased showing that two variants belonged to the same haplotype. Time efficiency is crucial in diagnosis as treatment (and its urgency) can greatly differ depending on etiology. Here, NanoCliD drastically reduced the diagnosis time compared to NGS short reads.

The last example is a duplication of exons 18 to 20 of *BRCA1* reported in NGS short reads. This rearrangement could lead to a mammary tumor. Only long reads sequencing was able to unravel a tandem duplication, classified as a pathogenic mutation. Adaptive sequencing and NanoCliD detected this event which was then reviewed manually using IGV, highlighting a tandem duplication leading to the pathogenic classification of the variant and thus improving the genetic counseling for the patient and her family.

4. Discussion

Nanopore adaptive sampling and NanoCliD improve clinical diagnosis by providing results for methylation profile, variant phasing, SVs and CNVs. Currently, only cases already analysed in short reads are passed through NanoCliD. Although NanoCliD is a diagnostic pipeline that allows to confirm/infirm a hypothesis established with other methods, it eventually could be used to detect SVs and variants without prior knowledge. For that purpose, the priority would be to reduce the noise generated by the variant callers. By using samples that have been sequenced with both long reads and short reads, we expect to be able to refine our filters to improve the tuning of NanoCliD's tools.

Allele phasing is also performed inside the pipeline. Fields in VCF such as PhaseSet ID can be used to highlight variants belonging to the same haplotype. Thus, haplotype reconstruction could be systematically provided in the final variant report. Another promising enhancement of NanoCliD could be promoter methylation analysis. Promoter methylation is involved in gene silencing in different cancer types [5]. Nanopolish generates the methylation levels of sufficiently covered CpG sites. This information could be used to investigate promoter methylation for a range of genes of interest.

References

1. Euskirchen P. and Bielle F. and Labreche K. *et al*. *Same-day genomic and epigenomic diagnosis of brain tumors using real-time nanopore sequencing.* Acta Neuropathologica, (134):691-703, 2017
2. Sakamoto Y. and Sereewattanawoot S. and Suzuki, A. *A new era of long-read sequencing for cancer genomics.* Journal of Human Genetics, (65):3-10,2020
3. Wang Y. and Zhao Y. and Bollas A. *et al. Nanopore sequencing technology, bioinformatics and applications.* Nature Biotechnology, (39):1348-1365, 2021
4. A. Payne, N. Holmes, T Clarke, et al. *Readfish enables targeted nanopore sequencing of gigabase-sized genomes.* Nature Biotechnology, (39):442–450, 2021
5. Kulis M. and Esteller M. *DNA methylation and cancer.* Advances in Genetics, (70):27-56, 2010

Jobim 2022 Poster
AskoR, an R package for easy RNA-Seq data analysis illustrated by the analysis of plant/pathogen/microbiote interactions

Susete Alves Carvalho*, Kévin Gazengel*, Sylvin Masanelli, Anthony Bretaudeau, Stéphanie Robin, Stéphanie Daval and Fabrice Legeai
INRAE - IGEPP, Domaine de la motte, 35650, Le Rheu, France
*contributed equally

Corresponding author: kevin.gazengel@inrae.fr

To make the process of transcriptomics data easier, and to guarantee the reproducibility of the analyses, we have developed AskoR, which is an R library to achieve a suite of statistical analyses and graphical outputs from gene expression data obtained by high-throughput sequencing (RNA-seq). From raw counts generated by mapping and counting tools, it allows to filter and normalize data, to check the consistency of samples, to perform differential expression tests, to execute GO-term enrichments, and to define co-expression clusters corresponding to expression patterns between experimental conditions. The edgeR package [1] was chosen for differential expression analyses, topGO [2] for GO-term enrichments and coseq [3,4] for co-expression clusters identification. Users can define a large number of parameters (about 60) as, for example, significance thresholds for statistical tests, algorithms for each tool or the generation of specific graphs. The tool has the advantage of being flexible : on the one hand, novices users can apply the default parameters defined on the basis of those commonly used, and on the other hand, experienced users can go further in the analyses by adapting these settings. All analysis steps automatically and quickly generate a large number of tables and figures in an output folder as summary tables for each step, expression heatmaps, volcanoplots, Venn diagrams or Upset graphs (less than 25 minutes for 16 transcriptomes and 20,000 genes). AskoR can be downloaded from Askomics GitHub (**https://github.com/askomics/askoR**) and used in the R environment or is directly accessible via a Galaxy portal on the GenOuest platform (https://galaxy.genouest.org). It is also planned to make it available on CRAN and allow its deployment on any Galaxy platform. Finally, AskoR produces outputs compatible with the AskOmics tool [5] developed to integrate complex data.

This tool is currently used for many different transcriptomics analyses. As an example, recently, it has been used for testing the transcriptomic fingerprint in *Brassica napus* and the pathogen responsible for clubroot *Plasmodiophora brassicae* when interacting with different levels of soil microbial diversity [6]. AskoR can therefore be used to analyze gene expression from RNA-seq experiments but can also be extrapolated to SmallRNA-seq or metagenomics data, as well as any other experiment that leads to the generation of a count table (excepted GO-term enrichment analysis).

References

1. Robinson MD. *et al.* edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics. 26(1):139-140. doi:10.1093/ bioinformatics/btp616, 2010.

2. Adrian A. and Rahnenfuhrer J. topGO: Enrichment Analysis for Gene Ontology. R package version 2.46.0, 2021.

3. Rau A. and Maugis-Rabusseau C. Transformation and model choice for co-expression analysis of RNA-seq data. Briefings in Bioinformatics, 19(3)-425-436, 2018.

4. Godichon-Baggioni A. *et al.* Clustering transformed compositional data using K-means, with applications in gene expression and bicycle sharing system data. Journal of Applied Statistics, doi:10.1080/02664763.2018.1454894, 2018.

5. Garnier X. *et al.* AskOmics: a user-friendly interface to Semantic Web technologies for integrating local datasets with reference resources. JOBIM, Nantes, France. pp.1. hal-02401750, 2019.

6. Daval S. *et al.* Soil microbiota influences clubroot disease by modulating Plasmodiophora brassicae and Brassica napus transcriptomes. Microb Biotechnol. 13(5):1648-1672. doi: 10.1111/1751-7915.13634. Epub 2020 Jul 19. PMID: 32686326; PMCID: PMC7415369, 2020.

ReMap 2022: a database of Human, Mouse, Drosophila and Arabidopsis regulatory regions from an integrative analysis of DNA-binding sequencing experiments

Fayrouz Hammal[1], Pierre de Langen [1], Aurélie Bergon[1], Fabrice Lopez[1], Benoit Ballester[1]

[1] Aix Marseille Univ, INSERM, TAGC, Marseille, France.

Corresponding Authors: fayrouz.hammal@inserm.fr, benoit.ballester@inserm.fr

1. Introduction

ReMap[1] is a resource of transcriptional regulators binding regions available in four different species: *Arabidopsis thaliana, Mus musculus, Drosophila melanogaster, Homo sapiens*. Our work aims to combine all publicly available ChIP-seq of transcriptional regulators (TRs) and create a catalog of manually curated and uniformly processed datasets. By its size and complexity, the ReMap catalogs allow a better understanding of the regulatory landscape. Currently we have annotated, curated, and uniformly processed a total of 8,103 ChIP-seq datasets for the human genome covering a total of 1,210 transcriptional regulators across 182 million peaks. The mouse catalog also contains a large amount of experiments with 5,503 datasets of 648 transcriptional regulators across 123 million of peaks.

2. Materials and Method

For the four species available in our ReMap catalog the same steps were used. The ChIP-seq datasets were retrieved from GEO[2] and ENCODE[3]. To create this regulatory atlas, we have manually curated and uniformly processed the ChIP-Seq datasets. Due to the heterogeneity of datasets, the pipeline assesses the quality of the datas and filters them accordingly . The pipeline used for the processing, filtering and control quality is available on github (https://github.com/remap-cisreg).

The four regulatory catalogs are available at https://remap.univ-amu.fr but are also browsable as native tracks in UCSC genome browser. Complex filtering features on targets or biotypes can be applied to improve visualization of ReMap peaks.

3. Conclusion

The ReMap database provides a high quality regulatory catalog in four species, by manually curating publicly available ChIP-seq datas, uniformly processing them and applying quality filters. Those datas are easily browseable on our website, and can also be visualized on UCSC genome browser.

Acknowledgements

PhD Fellowship to F.H. from the Provence-Alpes-Côte d'Azur Regional Council (Région SUD); Institut National de la Santé et de la Recherche Médicale (INSERM).

References

1. Hammal F., De Langen P., Bergon A., Lopez F., & Ballester B. (2022). ReMap 2022: a database of Human, Mouse, Drosophila and Arabidopsis regulatory regions from an integrative analysis of DNA-binding sequencing experiments. *Nucleic acids research*, *50*(D1), D316-D325
2. Barrett T., Wilhite S.E., Ledoux P., Evangelista C., Kim I.F., Tomashevsky M., Marshall K.A., Phillippy K.H., Sherman P.M., Holko M. et al. . NCBI GEO: archive for functional genomics data sets–update. *Nucleic Acids Res.* 2013; 41:D991–D995.
3. ENCODE Project Consortium An integrated encyclopedia of DNA elements in the human genome. *Nature*. 2012; 489:57–74.

Pipeline d'identification extensive de Variations Structurales dans du reséquençage nanopore de génome complet pour l'identification de mutations dans des maladies rares

Seydi Thimbo[1], Jean-François Deleuze[1], Eric Bonnet[1] et Claire Jubin[1]

[1] Centre National de Recherche en Génomique Humaine (CNRGH), Institut de Biologie François Jacob, CEA, Université Paris-Saclay, F-91057, Evry, France

Corresponding Author: claire.jubin@cnrgh.fr

L'identification de variations structurales (VSs) dans des données de reséquençage de génomes est un processus difficile. Ce type d'analyse qui s'est répandue avec l'avènement du séquençage massif et les NGS (New Generation Sequencing) s'est heurté à la présence de répétitions génomiques. Si on ne considère que le cas des SINEs (Small INterspersed repeat Elements) et LINEs (Long Interspersed Nuclear Elements), ces types d'éléments répétés représentent à eux seuls respectivement 13,1% et 20,4% du génome humain [1]. Que ce soient des duplications ou des régions de faible complexité, la fiabilité de la détection de VSs dans ces régions répétées est très faible, principalement en raison des faux positifs induits par les ambiguïtés d'alignement quand les régions répétées [2, 3] sont plus grandes que la taille des lectures de séquençage (50-300pb). L'usage des longues lectures (jusqu'à 1Mb) d'Oxford Nanopore Technologie (ONT) [4] permet de minimiser ce problème. Nous présentons ici un pipeline d'identification et d'expertise de VSs dans du reséquençage de génome complet dans la technologie ONT. Dans un projet de recherche de mutations responsables de maladies rares chez 5 patients, nous avons mis en évidence que les lectures de taille inférieure à 5kb étaient génératrices d'un bruit non négligeable pour l'identification de VSs dans une approche *mapping* [5]. Par ailleurs, cette approche étant contrainte par l'utilisation d'une séquence de référence, nous avons complété notre pipeline par une approche *assemblage de novo* [6] dans l'objectif de capter des VSs échappant à l'identification dans l'approche *mapping* de par leur trop grande différence avec la séquence de référence. Ces variants structuraux *stricto sensu* sont en partie des variants complexes, c'est-à-dire des compositions de variations simples que sont les délétions, duplications, inversion et translocations. Non seulement l'évaluation de la contiguïté des assemblages est très bonne (10.2Mb < N50 < 16Mb), mais en plus, l'analyse permet d'identifier des VSs spécifiques. Cela montre que l'utilisation de l'approche assemblage permet de progresser vers une identification exhaustive des VSs. Minimiser le nombre de faux négatifs est crucial, en particulier dans les projets de recherche de mutations responsables de maladies rares. Malheureusement, trouver de telles mutations n'est pas direct car de nombreux variants structuraux, principalement populationnels, existent entre le génome de n'importe quel individu et la séquence de référence. En dernière étape de notre pipeline, afin de cibler des VSs potentiellement responsables de la pathologie étudiée, nous générons des fichiers utilisateurs finaux dédiés à l'expertise des listes de VSs pour générer des listes de bons candidats selon différents critères.

References

1. Gregory, T.R., *Synergy between sequence and size in large-scale genomics.* Nat Rev Genet., 2005. **6**(9): p. 699-708. doi: 10.1038/nrg1674.
2. Lee, H. and M.C. Schatz, *Genomic dark matter: the reliability of short read mapping illustrated by the genome mappability score.* Bioinformatics, 2012. **28**(16): p. 2097-105.
3. van Belzen, I., et al., *Structural variant detection in cancer genomes: computational challenges and perspectives for precision oncology.* NPJ Precis Oncol, 2021. **5**(1): p. 15.
4. Wang, Y., et al., *Nanopore sequencing technology, bioinformatics and applications.* Nature Biotechnology, 2021. **39**(11): p. 1348-1365.
5. Sedlazeck, F.J., et al., *Accurate detection of complex structural variations using single-molecule sequencing.* Nat Methods., 2018. **15**(6): p. 461-468. doi: 10.1038/s41592-018-0001-7. Epub 2018 Apr 30.
6. Kolmogorov, M., et al., *Assembly of long, error-prone reads using repeat graphs.* Nat Biotechnol., 2019. **37**(5): p. 540-546. doi: 10.1038/s41587-019-0072-8. Epub 2019 Apr 1.

Harnessing gene expression profiling to infer the activation states of dendritic cell types, their dynamical relationships and their molecular regulation

Ammar Sabir CHEEMA [1], Thien-Phong VU MANH[1] and Marc DALOD[1]

[1] Aix Marseille Univ, CNRS, INSERM, CIML, Centre d'Immunologie de Marseille-Luminy, Turing Center for Living Systems, F-13288, Marseille, France.

Corresponding Author: cheema@ciml.univ-mrs.fr

Dendritic cells (DCs) are innate immune cells specialized in sensing danger signals and translating this information for the activation and functional polarization of effector immune cells, to fight infections and cancer. DCs are functionally plastic and this functional plasticity of DCs results from two key characteristics. First, DCs encompass distinct cell types that are specialized in different functions. Second, each DC type is itself plastic. It expresses a specific array of innate immune recognition receptors each able to sense different input signals. Depending on the precise combination of input signals received, a given DC type will undergo a particular activation state, to deliver a customized set of output signals activating effector immune cells in the way the best suited to fight the threat faced by the organism.

I developed a pipeline using previously existing tools for better identification of DC types and their activation states upon analysis of single cell RNA sequencing (scRNA-seq) data [1]. This pipeline allowed to efficiently separate type 1 conventional DCs (cDC1) from type 2 conventional DCs (cDC2), which is notoriously difficult to achieve at the single cell level when using only gene expression data [2]. It also helped understanding the effect of different pathophysiological conditions on the activation trajectory of cDC1.

Then this pipeline was used to better understand the role of cDC1 in a mouse model of melanoma. I focused on analyzing scRNA-seq data of tumor-associated cDC1 of WT or mutant genotypes. Coupled with wetlab experiments, this computational analysis contributed to determine how NF-κB-dependent IRF1 activation in cDC1 drives their antitumor activity [3].

Using my expertise in scRNA-seq data analysis combined with literature mining, I aim at generating a computational model of the molecular network regulating the production of type I interferon (IFN-I) by plasmacytoid DCs (pDCs) during a viral infection. pDCs are professional producers of IFN-I. This production is tightly regulated at single cell level (~10-15% of IFN-I-producing pDC), in space (spleen) and time (peak at 36 hrs). Failure of this tight regulation could lead to various autoimmune/inflammatory diseases, including Lupus erythematosus, psoriasis or Sjögren syndrome. Hence, better understanding the mechanisms regulating pDC IFN-I production could help treat these diseases. I am constructing regulatory networks between the molecules known to modulate pDC activation based on literature mining. I will complete them with candidate novel pathways inferred from gene regulatory network (GRN) analysis of our pDC scRNA-seq data from mice infected with murine cytomegalovirus [4]. Finally, I will design a logical model based on this regulatory network, using Ginsim or Bonesis, to infer the various activation states that pDC can adopt during the course of their activation for IFN-I production. This model will allow us predicting the outcome of define perturbation of the network, which my colleagues will test experimentally, to enable iterative cycles of model refinement to improve data fitting.

References

1. Ammar Sabir Cheema, Kaibo Duan, Marc Dalod, and Thien-Phong Vu Manh. Harnessing single cell RNA sequencing to identify dendritic cell types, characterize their biological states and infer their activation trajectory. *Methods Mol Biol*, In press.
2. Barbara Maier, Andrew M Leader, Steven T Chen, et al. A conserved dendritic-cell regulatory program limits antitumour immunity. *Nature*, (580):257-262, 2020.
3. Ghita Ghislat, Ammar Sabir Cheema, Elodie Baudoin, et al. NF-κB–dependent IRF1 activation programs cDC1 dendritic cells to drive antitumor immunity. *Science Immunology*, (6):eabg3570, 2021.
4. *Abdenour Abbas, *Thien-phong Vu Manh, Michael Valente, et al. The activation trajectory of plasmacytoid dendritic cells in vivo during a viral infection. *Nat Immunol*, (21):983-997, 2020.

scRNA-seq with Nanopore sequencing: benchmark of approaches based on hybrid sequencing

Ali HAMRAOUI[1,3], Catherine SENAMAUD-BEAUFORT[1], Stéphane LE CROM[2,1], Morgane THOMAS-CHOLLIER[1], Laurent JOURDREN[1] and Sophie LEMOINE[1]

[1] GenomiqueENS, Institut de Biologie de l'ENS (IBENS), Département de biologie, École normale supérieure, CNRS, INSERM, Université PSL, 75005 Paris, France

[2] Sorbonne Université, CNRS, Institut de Biologie Paris-Seine (IBPS), Laboratory of Computational and Quantitative Biology (LCQB), F-75005, Paris

[3] Master Bioinformatique Biologie-Informatique, Université Paris Cité

Corresponding Author: hamraoui@bio.ens.psl.eu

Long-read sequencing has recently been incorporated into single-cell RNA sequencing studies. Compared to the standard Illumina short-read library that primarily captures abundance information, Nanopore sequencing has several advantages : it can produce reads of several kb [1], it provides more precise transcriptome mapping/assembly and detects complex isoform variants.

In the 10xGenomics Nanopore data analysis, a key challenge is the relatively low sequencing accuracy (~ 95% per base) which makes it difficult to detect the cell barcodes and UMI information in each Nanopore read [1]. To deal with this issue, several bioinformatics approaches have been developed using hybrid sequencing to guide the allocation of Nanopore reads using Illumina data.

We performed a benchmark analysis of the available Single-Cell long-read hybrid sequencing based pipelines: Sicelore [2], Snuupy [3], Scnapbar [4], and Flames [5]. We first compared the performance of each pipeline for barcode-UMI detection and assignment, the ability of the pipeline to handle a large set of data sequenced on a PromethION, and the biological results at the gene level compared with Illumina results. We then selected the best tools to compare biological results at the transcript level.

Our results showed that Snuupy outperforms Sicelore, Flames and Scnapbar in terms of barcode-UMI assignment. The polyA-independant algorithm of Snuupy assigns around 35% more reads than Sicelore on a MinION dataset. This UMI count increase allows us to obtain biological results closer to Illumina results than sicelore at the gene level. Sicelore showed good performance processing a large set of data from PromethION where Snuupy could not achieve its process, suggesting it cannot be used without important modifications of its code. The results produced by Flames on the assignment part were not convincing but as it is designed for isoform detection, we are now testing its isoform detection module coupled with Sicelore and Snuupy.

Our aim is to select the best features in each pipeline to ultimately develop an optimal, scalable and reproducible pipeline to characterize transcript isoforms in single-cell data, using hybrid sequencing. We also foresee that the announced improvement of Nanopore sequencing accuracy may leverage the need of Illumina sequencing.

Acknowledgements

The IBENS genomics core facility 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. Full article: Single-cell transcriptomics in the context of long-read nanopore sequencing. https://www.tandfonline.com/doi/full/10.1080/13102818.2021.1988868.

2. High throughput error corrected Nanopore single cell transcriptome sequencing | Nature Communications. https://www.nature.com/articles/s41467-020-17800-6.

3. Long, Y. et al. FlsnRNA-seq: protoplasting-free full-length single-nucleus RNA profiling in plants. Genome Biol. 22, 66 (2021).

4. Wang, Q. et al. Single-cell transcriptome sequencing on the Nanopore platform with ScNapBar. RNA 27, 763–770 (2021).

5. Tian, L. et al. Comprehensive characterization of single-cell full-length isoforms in human and mouse with long-read sequencing. Genome Biol. 22, 310 (2021).

PANORAMA : comparative pangenomics tools to explore interspecies diversity of microbial genomes

Jérôme Arnoux[1], David Vallenet[1], Alexandra Calteau[1]

[1] LABGeM, Génomique Métabolique, CEA, Genoscope, Institut François Jacob, Université d'Évry, Université Paris-Saclay, CNRS

Corresponding Author: jarnoux@genoscope.cns.fr

1. Introduction

The last few years have seen the explosion of sequencing projects, leading to a considerable increase of genomes available in public databases. Comparative genomics approaches now use hundreds of genomes to analyze the diversity of a species. Many studies focus on the overall gene content of a species, the pangenome, to understand its evolution in terms of common and variable genes with regard to epidemiological or environmental data [1]. Nevertheless, the processing of such a mass of data imposes a paradigm shift in knowledge representation and in the algorithms used [2].

In this context, we developed PANORAMA, a bioinformatics toolbox, including new methodological developments for the comparative study of pangenomes.

2. Methods

PANORAMA benefits from methods for the reconstruction and analysis of pangenome graphs, thanks to the PPanGGOLiN software suite (https://github.com/labgem/PPanGGOLiN) [3], particularly for the identification of regions of genomic plasticity (RGPs) and their segmentation in conserved modules, with the panRGP [4] and panModule [5] methods. PANORAMA allows inter-pangenome analysis, relying on the exploration and comparison of RGPs and modules predicted in different species. Moreover, the tool proposes functional annotations of pangenomes families to provide an identification of functional systems, as secretion or conjugation system.

3. Results

PANORAMA offers rapid and easy to use comparative analyses of pangenomes on several thousand genomes of different species. PANORAMA will help to understand the adaptive potential of bacteria and, with the exploration of functional modules in different species, provide a better understanding of the evolutionary dynamics behind the metabolic diversity of microorganisms.

References

1. Golicz AA, Bayer PE, Bhalla PL, Batley J, Edwards D. Pangenomics Comes of Age: From Bacteria to Plant and Animal Applications. Trends Genet. 2020;36: 132–145.
2. Computational Pan-Genomics Consortium. Computational pan-genomics: status, promises and challenges. Brief Bioinform. 2016. doi:10.1093/bib/bbw089
3. Gautreau G, Bazin A, Gachet M, Planel R, Burlot L, Dubois M, et al. PPanGGOLiN: Depicting microbial diversity via a partitioned pangenome graph. PLoS Comput Biol. 2020;16: e1007732.
4. Bazin A, Gautreau G, Médigue C, Vallenet D, Calteau A. panRGP: a pangenome-based method to predict genomic islands and explore their diversity. Bioinformatics. 2020;36: i651–i658 doi:10.1093/bioinformatics/btaa792
5. Bazin, Adelme, Claudine Medigue, David Vallenet, et Alexandra Calteau. PanModule: Detecting Conserved Modules in the Variable Regions of a Pangenome Graph. bioRxiv, 2021. doi:10.1101/2021.12.06.471380.
6. Vallenet D, Calteau A, Dubois M, Amours P, Bazin A, Beuvin M, et al. MicroScope: an integrated platform for the annotation and exploration of microbial gene functions through genomic, pangenomic and metabolic comparative analysis. Nucleic Acids Res. 2020;48: D579–D589.

FROGSFUNC: Smart integration of PICRUSt2 software into FROGS pipeline

Vincent DARBOT[1], Moussa SAMB [1], Maria BERNARD[2], Olivier RUE[3] and Géraldine PASCAL[1]

[1] GenPhySE, Université de Toulouse, INRAE, ENVT, F-31326, Castanet Tolosan, France
[2] Univ. Paris-Saclay, INRAE, AgroParisTech, GABI, GIBBS/SIGENAE, F-78352, Jouy-en-Josas, France
[3] Univ. Paris-Saclay, INRAE, BioinfOmics, MIGALE bioinformatics facility, Jouy-en-Josas, France

Corresponding Author: vincent.darbot@inrae.fr

Metabarcoding is the large-scale taxonomic identification of complex environmental samples via analysis of DNA reads of one marker gene (16S, ITS, 18S, COI…).

The aim of metabarcoding analysis is to provide a table of abundance of OTUs/ASVs, as close as possible to the species, per sample as well as a descriptive statistical analysis of the composition of the targeted microbial population. The FROGS software offers such results [1, 2]. The different tools developed within FROGS allow users to process their data in command lines or via a user-friendly Galaxy interface and to obtain different graphical and descriptive outputs.

Metagenomics analysis allows the sequencing of all the DNA contained in a sample. It gives access to the functional potential of the species present in the environment [3]. Unlike metagenomics, metabarcoding does not provide these functional profiles of a microbial population, by being restricted to one marker gene. However, algorithms have been developed to bypass this restriction and obtain a prediction of the functional potential of a sample, at low cost. For this purpose, PICRUSt2 has been commonly used these last years [4]. Firstly, PICRUSt2 places the marker gene sequences (16S, ITS or 18S) of interest into its reference tree, that is used as the basis of functional predictions. After, it predicts number of marker and function copy number in each OTU/ASV. Then, for each sample, it calculates functions abundances and finally, pathway abundances are inferred, based on functional profile.

The plus of FROGS 4.0: additional outputs to the original PICRUSt2 tools that will guide users in the correct interpretation of their data and an ease of use thanks to the possibility of running these tools via Galaxy. Some "hidden" PICRUSt2 outputs are also exploited (reporting incongruence between taxonomic affiliations, NSTI threshold confidence indicator, decision support graphic to help choosing the NSTI threshold ...). In keeping with the FROGS philosophy, graphic outputs are also displayed to make the experience intuitive.

Funding

Vincent Darbot's work has been supported by RESALAB OUEST

References

[1] Escudié F, Auer L, Bernard M, Mariadassou M, Cauquil L, Vidal K, Maman S, Hernandez-Raquet G, Combes S, Pascal G. FROGS: Find, Rapidly, OTUs with Galaxy Solution. Bioinformatics. 2018 Apr 15;34(8):1287-1294. doi: 10.1093/bioinformatics/btx791. PMID: 29228191.

[2] Bernard M, Rué, O, Mariadassou M, and Pascal G; FROGS: a powerful tool to analyse the diversity of fungi with special management of internal transcribed spacers. Briefings in Bioinformatics. 2021, Nov. doi : 10.1093/bib/bbab318

[3] Thomas T, Gilbert J, Meyer F. Metagenomics - a guide from sampling to data analysis. Microb Inform Exp. 2012 Feb 9;2(1):3. doi: 10.1186/2042-5783-2-3. PMID: 22587947; PMCID: PMC3351745.

[4] Douglas GM, Maffei VJ, Zaneveld JR, Yurgel SN, Brown JR, Taylor CM, Huttenhower C, Langille MGI. PICRUSt2 for prediction of metagenome functions. Nat Biotechnol. 2020 Jun;38(6):685-688. doi: 10.1038/s41587-020-0548-6. PMID: 32483366; PMCID: PMC7365738.

metagWGS: a workflow to analyse short and long HiFi metagenomic reads

Maïna VIENNE[1], Jean MAINGUY[1], Pierre MARTIN[1,], Joanna FOURQUET[1,2], Céline NOIROT[1], Olivier BOUCHEZ[2], Adrien CASTINEL[2], Vincent DARBOT[3], Sylvie COMBES[3], Carole IAMPIETRO[2], Christine GASPIN[1], Denis MILAN[2], Cécile DONNADIEU[2], Géraldine PASCAL[3] and Claire HOEDE[1]

[1] INRAE, Université de Toulouse, UR875 MIAT, Bioinfomics, PF GenoToul Bioinfo, F-31326, Castanet-Tolosan, France (doi: 10.15454/1.5572369328961167E12)

[2] INRAE, GeT-PlaGe, Genotoul – INRAE – 31326 Castanet-Tolosan, France (doi: 10.15454/1.5572370921303193E12)

[3] GenPhySE, Université de Toulouse, INRAE, ENVT, F-31326, Castanet Tolosan, France

Corresponding Author: claire.hoede@inrae.fr

Whole DNA shotgun sequencing of environmental samples allows to study their taxonomic composition and their functional profiles. Recently, advances in sequencing technology and associated cost reductions allow the use of long high fidelity reads (HiFi) in metagenomics.

We are developing a complete, scalable, easy-to-use and reproducible workflow (with nextflow and Singularity), metagWGS, able to process short Illumina and long HiFi PacBio reads from shotgun metagenomics data. It provides (i) contig assemblies, (ii) syntactic and functional annotations of genes, (iii) taxonomic affiliations of reads and contigs, (iv) a counting table of reads per non redundant gene and (v) contigs binning to obtain Metagenome-Assembled Genomes*.
The workflow begins by preprocessing steps that clean raw data from adapters, low quality reads and the host reads. We control the quality of the reads with FastQC. The taxonomic classification of reads uses Kaiju in order to have a first overview of reads. The assembly is made by metaSPAdes or megahit for short reads and Hifiasm or metaFlye for long reads to generate contigs for each sample. This assembly can be realized per sample or as a co-assembly of several samples*.
Resulting contigs are annotated with Prokka. ORFs are clustered with CD-HIT using a 95% sequence identity cutoff to remove redundancy and generate a uniq gene catalog between samples. Reads are mapped back to contigs and featureCounts is used to count the reads overlapping annotated genes. The raw count table gathers the number of reads aligned on each gene for each sample. DIAMOND is used for the taxonomic affiliation of contigs versus nr database.
The binning step is being developed/implemented for both short reads and long reads assemblies.
At the end, MultiQC produces a report integrating all step global results.
We used metagWGS on several type of metagenomic data, we will show some results to illustrate the type of figures and tables obtained. We aim to provide training and support in the use of this workflow in the near future.

MetagWGS is available on https://forgemia.inra.fr/genotoul-bioinfo/metagwgs (GNU GPL License) with a complete and up to date documentation.

* in development at time of writing of this abstract.

Acknowledgements

SeqOccIn project funded by FEDER (Programme Opérationnel FEDER-FSE_Midi-Pyrénées et Garonne 2014-2020), ATB_Biofilm project funded by PNREST Anses, 2020/01/142, Antiselfish project funded by LabEx ECOFECT, Université de Lyon, ExpoMycoPig project funded by France Futur Elevage

References

We apologize, we don't have enough place to put the references.

Eco-evolutionary diversity of the global ocean microbiome across plankton size fraction

Nicolas Henry[1,2], Pierre E. Galand[2,3], Guillem Salazar[2,4], Marko Budinich[2,5], Charles Bachy[2,5], Erwan Delage[2,6], Mariana Câmara dos Reis[2,5], Hiroto Kaneko[2,7],Marinna Gaudin[2,6], *Tara Oceans Consortium,* Fabien Lombard[2,8], Hiroyuki Ogata[2,7], Silvia González Acinas[9], Patrick Wincker[2,10], Damien Eveillard[2,6], Shinichi Sunagawa[2,4], Samuel Chaffron[2,6], Christian Jeanthon[2,5] and Colomban De Vargas[2,5]

[1] CNRS - Sorbonne Université - Plateforme ABIMS - Station Biologique de Roscoff, Place Georges Teissier, 29680, Roscoff, France

[2] CNRS - Research Federation for the study of Global Ocean Systems Ecology and Evolution - FR2022/Tara Oceans GOSEE, Paris, France

[3] Sorbonne Université - CNRS - Laboratoire d'Ecogéochimie des Environnements Benthiques - LECOB - Observatoire Océanologique, 66650, Banyuls/Mer, France

[4] Department of Biology - Institute of Microbiology and Swiss Institute of Bioinformatics - ETH Zürich, Vladimir-Prelog-Weg 4, 8093, Zürich, Switzerland

[5] Sorbonne Université - CNRS - Laboratoire Adaptation et Diversité en Milieu Marin - Station Biologique de Roscoff, 29680, Roscoff, France

[6] Université de Nantes - CNRS - UMR 6004 - LS2N, 44000, Nantes, France

[7] Bioinformatics Center - Institute for Chemical Research - Kyoto University, Gokasho, Uji, Kyoto 611-0011, Japan

[8] Sorbonne Université - CNRS - UMR 7093 - Institut de la Mer de Villefranche-sur-Mer - Laboratoire d'Océanographie de Villefranche, 06230 Villefranche-sur-Mer, France

[9] Department of Marine Biology and Oceanography - Institute of Marine Sciences (ICM)–CSIC, Pg. Marítim de la Barceloneta, 37-49 Barcelona E08003, Spain

[10] Génomique Métabolique - Genoscope - Institut François Jacob - CEA - CNRS, Université Evry - Université Paris-Saclay, Evry, 91057 Paris, France

Corresponding Author: vargas@sb-roscoff.fr

Keywords: plankton, prokaryote, microbiome, metabarcoding, 16S

Abstract:

Prokaryotes (bacteria and archaea) are a critical component of the world ocean microbiome, and are classically divided between those having a true planktonic lifestyle ('free-living') and those associated with organic particles ('particle-attached'). Prokaryotes, however, thrive across all planktonic size fractions up to several centimetres, where they likely interact with eukaryotes in many, largely unknown ways, from parasitism to strict mutualism. We present here a unique global ocean metabarcoding dataset (16S rRNA gene) covering 1,131 water samples collected across 148 globally distributed *Tara* Oceans stations, 3 ocean depths, and 5 organismal size fractions from 0.2 µm to 2 mm. We describe highly diverse communities (>1.2 million prokaryotic ASVs) that we classified according to a newly defined plankton size index (PSI), and clustered them into evolutionary groups with consistent PSI using phylofactorization. We show that the prokaryotic communities from different size fractions have distinct patterns of diversity, evolution and environmental drivers. We further demonstrate that a significant proportion of the bacterioplankton, previously considered as 'particle-attached', actually represent microorganisms closely associated with eukaryotes in putative symbiotic relationships. This unprecedented dataset provides a first baseline description of the microbial plankton diversity in the world ocean and suggests the prevalence of interactions within marine microbial communities.

Construction of a reference genome catalog to decipher shared strains along an agrifood chain with shotgun metagenomic data

Solène PETY[1], Fiona BOTTIN[2], Sébastien THEIL[2], Céline DELBÈS[2], Panagiotis SAPOUNTZIS[3], Hélène CHIAPELLO[1], Pierre NICOLAS[1], Guillaume KON KAM KING[*1] and Anne-Laure ABRAHAM[*1]

[1] Université Paris-Saclay, INRAE, MaIAGE, 78350, Jouy-en-Josas, France

[2] Université Clermont-Auvergne, INRAE, VetAgro Sup, UMR545 Fromage, 15000, Aurillac, France

[3] Université Clermont Auvergne, INRAE, UMR 454 MEDIS, Clermont-Ferrand, France

(*) co-last authors

Corresponding Author: solene.pety@inrae.fr

Identifying fluxes of micro-organisms between successive compartments of an agrifood chain (soil, grass, litter, cow's feces and rumen, cheese) is important to understand and control cheese production. A first step to identify shared micro-organisms is to perform a taxonomical assignation at the species level. However, the sub-species / strain resolution is very relevant for the precise analysis of the assembly process of microbiota across habitats. In order to study strain fluxes, and take into account intra-species polymorphism, we choose an approach based on mapping metagenomic reads using the BWA-MEM tool [1] on a catalog of reference genomes to identify shared nucleotidic polymorphism across samples in our various ecosystems. The use of reference genomes instead of metagenomic assembled genomes allows capturing polymorphisms for more species than only the most abundant, and enables comparison across multiple datasets using a common reference.

Construction of a genome reference database is a key part of our analysis framework and must be tailored to the ecosystems under study. We will present the construction of a dedicated catalog based on the RefSeq [2] database with the addition of relevant genomes from different origins and projects to complete our database: metagenomic assembled genomes (MAGS) from previous experiments, and microbial genomes isolated from from cows' rumen and feces, and cheese. In particular, the species in the reference catalog must be different enough to avoid ambiguous mapping of the metagenomic reads, which requires aggregating similar genomes and choosing a representative for groups of aggregated species. Once the metagenomic reads mapped on a common reference, we will strive to reconstruct the various strains present in an ecosystem for the most abundant species by adapting existing methods (e.g. DESMAN [3]). Wherever coverage is insufficient to completely resolve strain genomes, we will use shared nucleotidic polymorphism between samples to compute similarity indices, based on Nei's distance [4], adapted to metagenomic samples.

References

1. Heng Li and Richard Durbin. Fast and accurate short read alignment with Burrows-Wheeler Transform. *Bioinformatics*, 25:1754-60, 2009.
2. Nuala A O'Leary, Mathew W Wright and *al.* Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. *Nucleic Acids Res.* 44(D1):D733-45, 2016.
3. Christopher Quince, Tom O. Delmont, Sébastien Raguideau, *and al.* DESMAN: a new tool for de novo extraction of strains from metagenomes. *Genome Biol* 18-181, 2017.
4. Masatoshi Nei. Estimation of average heterozygosity and genetic distance from a small number of individuals. *Genetics,* 89, 583 –590, 1978.

Full-length 16S rRNA gene MinION™ sequencing characterizing bacterial microbiota at species-level?

Valentine Gilbart[1,2], Rachel Bellone[1], Laurence Mousson[1], Jean-Philippe Martinet[1], Anna-Bella Failloux[1] and Catherine Dauga[1]
[1] Institut Pasteur, Arboviruses and Insect Vectors, 75015, Paris, France
[2] Master of Bioinformatics, University of Rouen Normandy, 76821, Mont-Saint-Aignan, France

Corresponding author: valentine.gilbart@pasteur.fr

Full-length 16S rRNA gene long-read sequencing offers the promise of species identification of bacteria. However, recent sequencing technologies such as MinION™ (Oxford Nanopore Technologies) exhibits high error rates, preventing high quality clustering and taxonomic assignation. No consensus methodology to analyze such data is yet established, but new nanopore-specific tools are emerging. In this work, we aim to optimize this process by comparing a classic OTU-based with a nanopore-specific approach for read clustering. For the purpose of assigning clusters to species-level, we also tested out different taxonomic assignation methods and databases.

The metagenomic data used came from a laboratory experiment where full 16S rRNA genes present in the carcass of *Aedes albopictus* mosquitoes were sequenced by 1D MinION™. We ran a typical OTU-based approach using mothur [1], preceded by Canu's [2] error-correction, that we compared to a 16S rRNA nanopore-specific approach with NanoCLUST [3]. Species of clusters were identified by BLAST, LCA and Naïve Bayesian Classifier, against the RefSeq and SILVA databases. For each reads, their cluster and taxonomic assignation with the different methods were compared.

NanoCLUST was more efficient in the clustering than OTU-based analysis. Species level assignation showed disagreement between classification approaches and databases, pointing out the need for further analysis to confidently identify species. Full-length 16S rRNA gene sequencing did allow species-level identification in some, but not all, cases. Out of the four biggest clusters, the best approach allowed to identify *Wolbachia* supergroup B and *Serratia marcescens* with accuracy, but failed for another *Wolbachia* cluster, and an *Asaia* cluster due to lack of variation in the 16S rRNA gene.

Altogether, our study highlighted the relevance of nanopore-specific tools to cluster and infer species assignation in long-read amplicon-based metagenomics studies, as well as the importance of the choice of the database and taxonomic assignation method.

References

[1] Patrick D. Schloss et al. Introducing mothur: Open-Source, Platform-Independent, Community-Supported Software for Describing and Comparing Microbial Communities. *Applied and Environmental Microbiology*, 75(23):7537–7541, December 2009. Publisher: American Society for Microbiology Section: METHODS.

[2] Sergey Koren, Brian P. Walenz, et al. Canu: scalable and accurate long-read assembly via adaptive k-mer weighting and repeat separation. *Genome Research*, 27(5):722–736, May 2017. Company, Distributor, Institution and Label: Cold Spring Harbor Laboratory Press Publisher: Cold Spring Harbor Lab.

[3] Héctor Rodríguez-Pérez, Laura Ciuffreda, and Carlos Flores. NanoCLUST: a species-level analysis of 16S rRNA nanopore sequencing data. *Bioinformatics*, 37(11):1600–1601, June 2021.

A metapangenomic approach for the association of prokaryotic genes to given phenotypic traits

Nicinthya PAJANISSAMY[1] and Guillaume GAUTREAU[1]

[1] Université Paris-Saclay, INRAE, MGP, 78350, Jouy-en-Josas, France

Corresponding Author: guillaume.gautreau@inrae.fr

1. Introduction

By collecting and comparing all the genomic sequences of a species, pangenomics studies focus on overall genomic content to understand species diversity 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) whereas the accessory part is crucial to understand the adaptive potential of prokaryotes and contains genomic regions that can be exchanged between strains by horizontal transfers.

Pangenomic approaches can be useful to understand microbial diversity, as the genetic varieties within a particular species can explain the phenotypic differences between groups of individuals. Indeed, comparisons of groups of genomes from individuals can raise links between the presence or absence of accessory genes and specific traits, which are helpful for example for diagnosis and prognosis. Moreover, advances in the methods of microbiota analysis based on shotgun sequencing can shed light upon these variations directly from samples comprising species that are hard to isolate.

Extending the PPanGGOLiN tool based on a pangenome graph framework, we propose to identify gene families contrasting between specific conditions. This identification can either be achieved by tagging the phenotypes of genomes belonging to the pangenome graph or directly by mapping reads from groups of samples over the pangenomes.

2. Method

Pangenomes are often modeled as binary presence/absence matrices, where the rows correspond to gene families and columns to the genome (1 in case of presence, 0 in case of absence). By specifying groups of genomes associated with specific conditions, it is then possible to identify contrasting families between these conditions through classical statistical approaches.

Moreover, by mapping metagenomic reads belonging to specific conditions, it is possible to infer the occurrence of genes of the pangenome in the samples in order to obtain a binary presence/absence matrice where the rows still correspond to families but where columns indicate the samples. Therefore, by applying the same statistical approaches as for the genomes, it is possible to determine the families explaining a particular contrast of phenotypes between samples.

3. Results

We first applied our approach to a dataset of genomes from antibiotic-resistant strains compared to sensitive ones in order to identify the families associated with this resistance. We also test our metapangenomics approach on fecal microbiota of patients with diseases compared to healthy controls using shotgun metagenomics. Thanks to the PPanGGOLiN features, it is then possible to explore the genomic contexts of these highlighted families through the visualization of the pangenome graph.

References

1. Guillaume GAUTREAU *et al*. PPanGGOLiN: Depicting microbial diversity via a partitioned pangenome graph *PLOS Computational biology*, 2022.

Keywords – pangenomics, microbiome, disease prediction, shotgun metagenomics, GWAS, pangenome graph

Deciphering molecular mechanisms governing malignant transformation in Neurofibromatosis Type 1 (NF1) from single cell transcriptomic

Audrey ONFROY[1,2], Fanny COULPIER[1], Katarzyna RADOMSKA[1], Laure LECERF[1], Denis THIEFFRY[2] and Piotr TOPILKO[1]

[1] Institut Mondor de Recherche Biomédicale, Inserm U995 Team 9, Créteil, France.
[2] PSL Université, IBENS - UMR ENS - CNRS 8197 - INSERM 1024, Paris, France

Corresponding Author: audrey.onfroy@inserm.fr

Context

Malignant peripheral nerve sheath tumors (MPNSTs) are aggressive sarcomas with no effective therapy to date. They mainly occur in patients with Neurofibromatosis Type 1 (NF1), a genetic disorder caused by loss of functions of Nf1 gene in Schwann cells. MPNSTs originate from benign nerve sheath tumors called plexiform neurofibromas (pNFs). Progression of pNFs into malignancy is preceded by a transient state, called dysplastic NF (dNF). While dNFs are present in about half of NF1 patients, they rarely transform into malignancy. Presumably, the capacity of dNFs to become malignant is dictated by the acquisition of genetic and epigenetic alterations. Our team has developed a mouse model (Nf1-KO) that recapitulates the development of pNFs and their malignant transformation. My project aims to apply multi-omics approach on tumors from Nf1-KO mice and from NF1 patients to decipher molecular events underlying malignant transformation. Currently, single-cell RNA sequencing (scRNA-Seq) datasets are available.

Cell type annotation at a single cell level

NFs and MPNSTs are made from a variety of cell types, whose heterogeneity reaches a maximum during dysplastic state. To study the cross-talks between these populations, identifying them is crucial. To our knowledge, there is no specific signature in common public databases to identify NF1 tumor cells. Previous results of our lab led to the identification of a few markers for several cell types forming the tumor. Moreover, all Nf1 mutant cells in the Nf1-KO mouse model can be easily identifies thanks to expression of the Tomato fluorescent reporter.

To strengthen the signatures associated with each tumor forming population, we used our mouse datasets to extend our pre-existing lists of markers. We further define negative signatures as the union of all positive markers for all other cell populations. On this basis, we attribute two expression scores to each cell and each cell type, using Seurat's AddModuleScore [1], for positive and negative markers, respectively. The difference between these two scores then reflects the possibility of a cell to belong to the corresponding cell type. Hence, each cell can be annotated with the cell type corresponding to the highest score, yet keeping track of potential overlapping profiles.

This method is performed at a single cell level, and does not include any clustering step. The annotation runs within hundred seconds, on 50 000 cells issued from Nf1-KO nerve sheath tumors. We checked known markers expression to validate the automatic annotation of stromal cells, and the expression level of Tomato for tumor cells. All immune cells were further annotated thanks to the Tumor Immune Cell Atlas [2], with consistent results. The algorithm seems promising on NF1 patients datasets, although no specific markers are known to identify tumor cells to date.

Prospects

Our next goal is to identify tumor subpopulations and assess their correlation with tumor malignancy. This should allow to identify actionable targets involved in malignant transformation.

References

1 Stuart T., Butler A. et al. , Cell (2019). DOI:10.1016/j.cell.2019.05.031
2 Nieto P., Elosua-Bayes M., Trincado J.L. *et al.*, Genome Research (2021). DOI:10.1101/gr.273300.120

Towards omics-based distribution modelling of marine plankton associations at global scale

Marinna GAUDIN[1], Damien EVEILLARD[1] and Samuel CHAFFRON[1]
Nantes Université, École Centrale Nantes, CNRS, LS2N, UMR 6004, F-44000 Nantes, France

Corresponding authors: marinna.gaudin@univ-nantes.fr, samuel.chaffron@univ-nantes.fr

Marine plankton communities are composed of microorganisms in perpetual interactions, which play an essential role in the functioning of our biosphere as they contribute to the regulation of major biogeochemical cycles, support half of the primary production, and form the basis of the food chain sustaining the global marine trophic network [1]. Increasing acidification and warming of the surface ocean have been associated with significant transformations of these plankton communities, impacting primary production, diversity and species biogeography [2]. The RCP8.5 scenario (high emission future without effective climate change mitigation policies) of the 2013 IPCC report [3] predicts a temperature increase of 6°C by the end of the century with a further decrease in pH of 0.3 to 0.4 units [4]. Such climate transformations could compromise the support of environmental functions by plankton communities, ultimately affecting the entire biosphere [5].

Predicting and characterising the nature and magnitude of future plankton ecological and functional responses has motivated the development of Species Distribution Modelling (SDM), a statistical framework that aims to model and predict the biogeography of plankton species in relation to their environmental context. SDM usually only takes into account physico-chemical parameters, neglecting the effects of biotic interactions between species. However, species interactions are essential factors influencing the dynamics of plankton communities, and may respond differently to environmental stressors than individual species [6]. The rise of meta-omics techniques provides a high-resolution framework to measure the diversity and abundance of plankton species directly from environmental samples, which led to the development of computational techniques aimed at detecting significant associations between plankton species, with the general assumption they might indicate potential interactions [7].

Here, we propose a computational framework to improve the modelling and prediction of plankton associations biogeography from meta-omics data, with the goal to identify oceanic regions and associated environmental parameters that will likely impact plankton community stability and resilience. We show that our modelling framework provides better predictive power for plankton biogeography as compared to classical SDM. Our computational framework connects ecological and climate modeling by combining species associations niche modeling with network analyses for predicting ecosystem-scale vulnerabilities to environmental change, with the goal to improve predictions of climate change effects on marine plankton resilience and impacts on associated ecosystem services.

References

[1] Paul Frémont, Marion Gehlen, Mathieu Vrac, Jade Leconte, Tom O. Delmont, Patrick Wincker, Daniele Iudicone, and Olivier Jaillon. Restructuring of genomic provinces of surface ocean plankton under climate change. *Nature Climate Change*, 12:393–401, 2022.

[2] Martin Edwards, Eileen Bresnan, Kathryn Cook, Mike Heath, Pierre Helaouet, Christopher Lynam, Robin Raine, and Claire Widdicombe. Impacts of climate change on plankton. *MCCIP Science Review*, 4, 2013.

[3] Stocker, T.F., D. Qin, G.-K. Plattner, M. Tignor, S.K. Allen, J. Boschung, A. Nauels, Y. Xia, and V. Bex. Intergovernmental panel on climate change : The physical science basis. *Cambridge University Press*, 2013.

[4] Richard Feely, Christopher Sabine, Kitack Lee, Will Berelson, Joanie Kleypas, Victoria Fabry, and Frank Millero. Impact of anthropogenic co2 on the caco3 system in the oceans. *Science*, 305:362–366, 2004.

[5] Benedetti, F., Vogt, M., Elizondo, and U.H. et al. Major restructuring of marine plankton assemblages under global warming. *Nature Communications*, 12(5226), 2021.

[6] Alfonso Valiente-Banuet et al. Beyond species loss: the extinction of ecological interactions in a changing world. *Functional Ecology*, 29:299–307, 2015.

[7] Morales-Castilla, Matias, Gravel, and Araujo. Inferring biotic interactions from proxies. *Trends in Ecology Evolution*, 30:347–356, 2015.

Characterising competition and cooperation potentials in microbial communities using discrete models of metabolism

Maxime Lecomte[1,2], David Sherman[2], Hélène Falentin[1] and Clémence Frioux[2]
[1] INRAe UMR STLO, 65 rue de Saint-Brieuc, 35042, Rennes, France
[2] Inria University of Bordeaux, 200 avenue de la Vieille Tour, 33405, Talence, France

Corresponding author: maxime.lecomte@inria.fr

Deciphering to what extent competition and cooperation are likely to occur in microbial communities result in non naive and time consuming steps. Modeling the metabolism, with accurate prediction of exchanged metabolite concentrations based on numerical simulations, need curated genome-scale metabolic networks and is hardly scalable. To identify the cooperation and competition potential of a community, the prediction accuracy can also be reduced to whether a reaction can produce a compound or not, reducing the computational cost. In [1,2], the competition and cooperation potential is based on pairwise analysis using scores combined with numerical methods such as Flux Balanced Analysis (FBA) [3]. While the numerical optimisation problem constrains to pairwise analysis, boolean abstraction, such in [4], decribes the metabolic potential of a possibly large community. Following this lead, we propose a method that enriches the description of a natural community.
We first described the metabolic potential of community in its environment, as well as putative accumulated metabolites and limiting substrates using metabolic networks and Answer Set Programming (ASP) [5]. From this description, we established metrics in order to obtain cooperation and competition scores. Regarding cooperation, we defined three sub-metrics focusing on the number of producible metabolites, activated reactions and possibly exchanged metabolites. Concerning competition, we focused on consumed metabolites to model the competition for resources.
In order to test our scores, we built random communities starting from a set of 1520 genomes of cultivable bacterial species isolated in the human gut microbiota [1]. We randomly designed 50 communities ranging from 5 to 200 species without replacement and analysed the distribution of the scores. We then tested the relevance of our metrics against data from [6], which performs *in vitro* co-growth experiments. Our first results show correlations between our predictions and observations in [6]. Overall, our method aims at facilitating the metabolic characterization of microbial communities and the design of customized communities for which numerical methods are not easily applicable.

Acknowledgements

We acknowledge the GenOuest bioinformatics core facility [2] for providing the computing infrastructure.

References

[1] Shiri Freilich, Raphy Zarecki, Omer Eilam, Ella Shtifman Segal, Christopher S. Henry, Martin Kupiec, Uri Gophna, Roded Sharan, and Eytan Ruppin. Competitive and cooperative metabolic interactions in bacterial communities. *Nature Communications*, 2(1), 2011.

[2] Aleksej Zelezniak, Sergej Andrejev, Olga Ponomarova, Daniel R. Mende, Peer Bork, and Kiran Raosaheb Patil. Metabolic dependencies drive species co-occurrence in diverse microbial communities. *Proceedings of the National Academy of Sciences of the United States of America*, 2015.

[3] Jeffrey D. Orth, Ines Thiele, and Bernhard O. Palsson. What is flux balance analysis?, 2010.

[4] Arnaud Belcour, Clémence Frioux, Méziane Aite, Anthony Bretaudeau, Falk Hildebrand, and Anne Siegel. Metage2metabo, microbiota-scale metabolic complementarity for the identi1cation of key species. *eLife*, 2020.

[5] Vladimir Lifschitz. What is answer set programming? *Proceedings of the National Conference on Artificial Intelligence*, 3:1594–1597, 2008.

[6] Anna S. Weiss, Anna G. Burrichter, Abilash Chakravarthy Durai Raj, Alexandra von Strempel, Chen Meng, Karin Kleigrewe, Philipp C. Münch, Luis Rössler, Claudia Huber, Wolfgang Eisenreich, Lara M. Jochum, Stephanie Göing, Kirsten Jung, Chiara Lincetto, Johannes Hübner, Georgios Marinos, Johannes Zimmermann, Christoph Kaleta, Alvaro Sanchez, and Bärbel Stecher. In vitro interaction network of a synthetic gut bacterial community. *ISME Journal*, 16(4):1095–1109, 2022.

1. http://dx.doi.org/10.1038/s41587-018-0008-8
2. https://www.genouest.org

Analyzing and modelling functions carried by key species in minimal microbial communities

Chabname GHASSEMI NEDJAD[1] and Clémence FRIOUX[1]

[1] Inria Univ. Bordeaux, INRAE, 33400 Talence, France

Corresponding Author: chabname.ghassemi-nedjad@inria.fr

Microbial communities play an important role in various environments, and as a result also is the gut microbiota, where host-microbial and microbe-microbe interactions have been highlighted. Numerous studies throughout the last decade illustrated these symbioses and their impact on the health of the host. By using Genome-Scale Metabolic Networks (GSMNs), one can build models describing the ecosystem at the level of metabolism and make hypotheses on the organization of the associated microbial communities [1]. With the improvement of GSMN reconstruction toolboxes, such models can be built for thousands of organisms, requiring scalable and robust metabolism abstractions.

Metage2Metabo (M2M) [2] is a tool screening all metabolism of a group of bacteria to detect key species and select minimal communities able to produce target metabolisms. M2M uses a discrete model to simulate the concept of producibility in metabolic networks and solve combinatorial optimization problems with ASP [3]. M2M was used on a set of more than 1,500 intestinal bacterial genomes [4]. With this study it was possible to design hundreds of thousands equivalent minimal communities and select key species i.e. bacteria predicted on one or more of these communities. The connections between key species were represented in power graphs [2], helping us to visualize equivalence groups of species, often grouped by phylum. These models are a first step towards the understanding of interactions between organisms, through the composition of minimal communities, but the metabolic mechanisms explaining these associations of key species were lacking.

By using the same programming paradigm as M2M, we designed new models to highlight the metabolism in equivalence groups of bacteria. We studied these groups through the lens of the functions they carry, the compounds they produce and reactions they can activate. We made comparisons between groups and also with bacteria that are not key species and were thus left out of the minimal communities. We applied our models to minimal communities for 5 groups of target metabolites predicted with M2M [2]. The results show the role of groups of GSMNs in unlocking functions for GSMNs of other groups, which groups are able to produce targets and the specificities around these productions. More generally our work proposes the possible interactions between members of minimal communities.

Our study succeeded at suggesting explainable models of previous combinatorial optimization problems results. With an expressive programming paradigm and a discrete model, we are able to propose metabolic mechanisms for a better understanding of microbial communities.

Acknowledgements

Experiments presented in this paper were carried out using the PlaFRIM experimental testbed, supported by Inria, CNRS (LABRI and IMB), Université de Bordeaux, Bordeaux INP and Conseil Régional d'Aquitaine (see https://www.plafrim.fr).

References

1. Frioux C, Singh D, Korcsmaros T, Hildebrand F. From bag-of-genes to bag-of-genomes: metabolic modelling of communities in the era of metagenome-assembled genomes. *Computational and Structural Biotechnology Journal* 18 (2020) 1722–1734
2. Belcour A, Frioux C, Aite M, Hildebrand F, Siegel A. Metabolic Complementarity for the Identification of Key Species. *ELife*, *9*, 1–33.
3. M. Gebser, A. König, T. Schaub, S. Thiele and P. Veber. The BioASP Library: ASP Solutions for Systems Biology. *2010 22nd IEEE International Conference on Tools with Artificial Intelligence*, 2010, pp. 383-389.
4. Y. Zou, W. Xue, G. Luo, Z. Deng, P. Qin, R. Guo, H. Sun, Y. Xia, S. Liang, Y. Dai1, D. Wan, R. Jiang, L. Su, Q. Feng, Z. Jie, T. Guo, Z. Xia1, C. Liu1, J. Yu1, Y. Lin, S. Tang, G. Huo, X. Xu, Y. Hou, X. Liu, J. Wang, H. Yang, K. Kristiansen, J. Li, H. Jia and L. Xiao. 1,520 reference genomes from cultivated human gut bacteria enable functional microbiome analyses. *Nat. Biotechnol*. 2019 372, vol. 37, no. 2, pp. 179–185, Feb. 2019.

Construction of ASVs networks to monitor the temporal dynamics of bacterial communities - Application to public datasets on vegetable fermentations

Romane JUNKER[1], Victoria CHUAT[2], Florence VALENCE[2], Michel-Yves MISTOU[1], Julie AUBERT[3], Stéphane CHAILLOU[4] and Hélène CHIAPELLO[1]

[1] Université Paris-Saclay, INRAE, MaIAGE, 78350, Jouy-en-Josas, France

[2] INRAE, Agrocampus Ouest, STLO, 35042, Rennes, France

[3] Université Paris-Saclay, AgroParisTech, INRAE, UMR MIA-Paris, 75005, Paris, France

[4] Université Paris-Saclay, INRAE, MICALIS, 78350, Jouy-en-Josas, France

Corresponding Author: romane.junker@inrae.fr

Using network approaches to characterize a set of microbiota samples and identify bacterial communities' evolution is a difficult task that requires dedicated approaches. Our project consists in designing generic bioinformatics tools to describe and represent bacterial communities and temporal dynamics in datasets of metabarcoding samples. In this work, we present a workflow to construct and represent bacterial communities and consortia and apply it to public lacto-fermented vegetables samples. Fermented foods have gained interest over the last five years, driven by the public's desire to eat healthy, minimally processed foods with low energy consumption for preparation and storage. If some key species of vegetable fermentations have been identified, there is, to our knowledge, no detailed description of the succession of bacterial communities and their interactions. From an Open Science perspective, we are interested in reusing published datasets. We searched for public 16S metabarcoding datasets related to fermented vegetables with a sufficient number of samples, clear and detailed metadata, and an associated publication in public databases (ENA, MGnify [1]).

Only the Wuyts et al. dataset on carrot juice fermentation [2] met these criteria. Its interest lies in the diverse samples present, from laboratory and participatory science experiments. An approach based on association networks between amplicon sequence variants (ASVs) was adopted: the networks were constructed from the count tables obtained (using the dada2 pipeline [3]) after filtering the ASVs on their abundance and prevalence thanks to the phyloseq R package [4]. Metrics-based methods were used to detect co-presence (Jaccard distance) and co-abundance (Pearson and Spearman correlation on relative abundances, proportionality on clr-transformed abundances with the ALDEx2 R package [5]) between ASVs. The following R packages were used to visualize the networks: ggiraph, ggraph, and igraph.

The networks made it possible to represent the ASVs diversity and show the bacterial communities succeeding each other during the different fermentation phases. We will present the workflow and discuss networks obtained on the Wuyts et al. dataset. Based on robust and standard R libraries, this strategy will allow the integration of new 16S metabarcoding datasets from vegetable fermentation.

References

1. Mitchell, Alex L., et al. "MGnify: the microbiome analysis resource in 2020." *Nucleic acids research* 48.D1 (2020): D570-D578.
2. Wuyts, Sander, et al. "Carrot juice fermentations as man-made microbial ecosystems dominated by lactic acid bacteria." *Applied and environmental microbiology* 84.12 (2018): e00134-18.
3. Callahan, Benjamin J., et al. "DADA2: High-resolution sample inference from Illumina amplicon data." *Nature methods* 13.7 (2016): 581-583.
4. McMurdie, Paul J., and Susan Holmes. "phyloseq: an R package for reproducible interactive analysis and graphics of microbiome census data." *PloS one* 8.4 (2013): e61217.
5. Gloor, Greg. "ALDEx2: ANOVA-Like Differential Expression tool for compositional data." ALDEX manual modular 20 (2015): 1-11.

An eucaryote-friendly set of python scripts for multi-sample shotgun metagenomics and its associated Shiny application for microbiota exploration

Fiona Bottin[1] and Sébastien Theil[1]
[1]Université Clermont-Auvergne, INRAE, VetAgro Sup, UMR545 Fromage, 20 côte de Reyne, F-15000, Aurillac, France

Corresponding author: fiona.bottin@inrae.fr

Shotgun metagenomic sequencing is a highly sensitive method to characterize microbiota diversity and functional profiles in an ecosystem. However, the drawbacks of its precision are the computer resources required and the large amount of resulting data, particularly on high throughput sample analyses. As a result, eukaryotes are often discarded during processes despite there roles in many microbial communities.

Considering these issues, we developed python tools for eukaryote-friendly multi-sample shotgun metagenomics analyses. They are based on pre-existing tools selected for their performance, their IT resource and their consideration of eukaryotic microorganisms. To optimize computational time of processes, they are parallelized on samples or on splited samples for the most time-consuming tasks. The scripts have been designed to be reproducible, adaptable and independent, but can easily be used as a workflow thanks to a yaml file that links inputs and outputs between successive steps. Currently, 6 scripts have been developed to propose a processing method for the following common steps of shotgun metagenomic analysis : taxonomy profiling of sample, advanced reads quality filtering, assembly, binning, taxonomic annotation and structural gene annotation of assembled contigs.
The suite of tools is available on https://forgemia.inra.fr/sebastien.theil/metag-tools, and development is still in progress.

Our laboratory focuses its studies on microbial and gene transfers between ecosystems inside a farm system of cheese production. In order to understand the food chain microbiota and its response to stress (drought, feed change...), we are developing a Shiny application to visualize ecosystem compositions and identify shared or specific elements between them. For an efficient querying, whole data must previously be structured and stored in a DuckDB[1] Database. Once uploaded, user can either explore the whole dataset or sub-select samples depending on metadata (specific environment types, collection dates,...) or on a specific taxa for the further analyses. The resulting dataset is then used to generate taxonomic plots from contigs, genes or bins. This module characterize microorganisms composition of samples or any metadata group according to a chosen taxonomic rank. A pair-wise comparison module allows to outline common or specific sequences between two groups or samples, based on a log2ratio value which describe a normalized differential abundance of sequencing depth. This analysis allows to explore taxonomy and functions of common or specific fraction of the selected dataset. A term enrichment analysis, enables to highlights KEGG/GO terms that are over or under represented when comparing, for instance, the overrepresented sequences versus the whole dataset. To overview the whole pair-wise results, they will be summarized on a network of shared terms between each samples or metadata groups. This last step enable to compare the microbiota composition between all studied ecosystems. Once functional, the application will be available on https://forgemia.inra.fr/sebastien.theil/exploremetag.

References

[1] Hannes Mühleisen Mark Raasveldt. Duckdb: an embeddable analytical database. *Association for Computing Machinery*, page 1981–1984, 2019.

LibProtein: a rapid and versatile annotation library for protein post-translational annotations

Hamady BA[1], and Stéphane Téletchéa[1]

[1] Nantes Université, CNRS, US2B, UMR6286, 2 chemin de la houssinière, F-44000, Nantes, France

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

1 Introduction

Protein are complex biological molecular entities defined primarily by their amino acids sequence. Upon or after ribosomal translation, the messenger RNA is transformed into a polypeptide chain, often these amino acids may further be processed into more complex entities with the addition of functional groups such as carbohydrates, pyrophosphates or other small chemical entities. [1], [2]

Since these co- or post-translational modifications are diverse and spread into multiple databases it is difficult to assemble these data, and therefore even more complex to reuse them for bioinformatics studies. The aim of this project is to ease protein PTM retrieval and annotation, and to provide usage examples.

2 Implementation and user cases

LibProtein is a C++ library allowing the retrieval and annotation of any given protein with modifications available first in uniprot. It contains an internal representation of PTM as a 3-letter code. This library also contains annotations from SCOP, CATH and PFAM to enrich sequences with functional domains descriptions.

The library implementation is flexible enough to envision the incorporation of experimental or predictions of PTMs.[3], [4] These additional annotations will provide a more complete overview of a given protein life cycle.

The library allows to enrich structure analysis in PyMol by the addition of annotations tags into the protein sequence (PTM, families and categories), to provide data for machine learning development, and to score protein structure models often lacking PTM predictions. [5], [6]

Acknowledgements

Hamady Ba's internship is funded by the TROPIC project which received financial support from the ANR grant "Programme d'investissements d'Avenir" (ANR-16-IDEX-0007), from Région Pays de La Loire and from Nantes Métropole.

References

[1] D. Tsikas, « Post-translational modifications (PTM): analytical approaches, signaling, physiology and pathophysiology—part I », *Amino Acids*, vol. 53, nº 4, p. 485-487, avr. 2021, doi: 10.1007/s00726-021-02984-y.

[2] A. G. de Brevern et J. Rebehmed, « Current status of PTMs structural databases: applications, limitations and prospects », *Amino Acids*, janv. 2022, doi: 10.1007/s00726-021-03119-z.

[3] M. Audagnotto et M. Dal Peraro, « Protein post-translational modifications: In silico prediction tools and molecular modeling », *Comput. Struct. Biotechnol. J.*, vol. 15, p. 307-319, janv. 2017, doi: 10.1016/j.csbj.2017.03.004.

[4] Y. Deng, Y. Fu, H. Zhang, X. Liu, et Z. Liu, « Protein Post-translational Modification Site Prediction using Deep Learning », *Procedia Comput. Sci.*, vol. 198, p. 480-485, 2022, doi: 10.1016/j.procs.2021.12.273.

[5] J. Jumper et D. Hassabis, « Protein structure predictions to atomic accuracy with AlphaFold », *Nat. Methods*, vol. 19, nº 1, p. 11-12, janv. 2022, doi: 10.1038/s41592-021-01362-6.

[6] M. Baek *et al.*, « Accurate prediction of protein structures and interactions using a three-track neural network », *Science*, vol. 373, nº 6557, p. 871-876, août 2021, doi: 10.1126/science.abj8754.

An integrative bioinformatics approach to explore the biodiversity of enzyme families

Eddy Elisée[1], Mark Stam[1], Raphaël Méheust[1], Carine Vergne-Vaxelaire[1], David Vallenet[1]

[1] Génomique Métabolique, CEA, Genoscope, Institut François Jacob, Université d'Évry, Université Paris-Saclay, CNRS

Corresponding Author: eelisee@genoscope.cns.fr

1. Introduction

Metagenomics data represent a largely untapped and continuously growing pool of new sequences coming from various worldwide biotopes (soil, human gut, oceans…). This biodiversity, up to several billions of sequences, may be exploited to answer multiple scientific challenges. For instance, biocatalysis (i.e. catalysis with enzymes) needs new relevant biocatalysts with various activities to take up the energy transition challenge and replace some polluting synthesis steps. Hence, bioinformatics approaches to efficiently identify the targeted enzymatic activity from large metagenomic resources are needed.

Through the MODAMDH project (ANR JCJC), we focused on one of the key biocatalysts named amine dehydrogenases (AmDHs) which enable the access to amines that are important entities in the chemical industry [1-2]. To do so, we applied a sequence- and structure-based bioinformatics approach to widen the landscape of protein sequences catalyzing reductive amination by searching for remote homologs and active site analogs. In the context of the ALADIN project (ESR / EquipEx+), we want to generalize this approach and develop workflows that could be applied to any enzyme family.

2. Methods

Publicly available and in-house metagenomics databases (>2.5 billion protein sequences) were screened using HMMER software and the SUPERFAMILY database. Structural modeling and active site classification were performed by the ASMC software [3]. Remote homologs were recovered by HMM-HMM comparisons with HHblits software. Active site analogs were searched by screening catalophores (i.e. minimal active site topologies) using the YASARA software. Phylogeny of protein families was done with IQ-TREE program.

3. Results

The AmDH family was first enriched with new metagenomic sequences before being classified into subfamilies using an active site classification and a phylogeny. Besides, we generated a pool of NAD(P)-binding protein sequences from which we found, using HMM-HMM comparisons, new AmDH distant homologs. In contrast, no active site analog has yet been found for the AmDH family.

Through the ALADIN project, we will extend this strategy to other enzymatic activities by designing generic workflows and applying them first to explore the diversity of the aforementioned NAD(P)-binding protein families.

Acknowledgements

This study was supported by the contracts from the MODAMDH (ANR-19-CE07-0007, ANR JCJC) and ALADIN (IA-21-ESRE-0021, ESR / EquipEx+) projects.

References

1. Ducrot L, Bennett M, Grogan G, Vergne-Vaxelaire C. NADP(H)-Dependent Enzymes for Reductive Amination: Active Site Description and Carbonyl-Containing Compound Spectrum. Adv. Synth. Catal. 2020, 363, 328-351. doi:10.1002/adsc.202000870.
2. Mayol O, Bastard K, Beloti L *et al*. A family of native amine dehydrogenases for the asymmetric reductive amination of ketones. Nat. Catal. 2019, 2, 324–333. doi:10.1038/s41929-019-0249-z.
3. de Melo-Minardi RC, Bastard K, Artiguenave F. Identification of Subfamily-specific Sites based on Active Sites Modeling and Clustering. Bioinformatics 2010, 26(24), 3075-3082. doi:10.1093/bioinformatics/btq595.

Towards molecular understanding of the UbiJ-UbiK$_2$ protein complex by multiscale molecular modelling studies

Romain LAUNAY[1], Elin TEPPA[1], Carla MARTINS[1], Sophie ABBY[2], Fabien PIERREL[2], Isabelle ANDRE*,[1] and Jérémy ESQUE*,[1]

[1] Toulouse Biotechnology Institute, TBI, Université de Toulouse CNRS, INRAE, INSA, Toulouse, France. 135, avenue de Rangueil, F-31077 Toulouse Cedex 04, France

[2] TIMC, Université Grenoble Alpes, CNRS, CHU Grenoble Alpes, Grenoble INP, Grenoble, France

Corresponding Authors: isabelle.andre@insa-toulouse.fr, jeremy.esque@insa-toulouse.fr

Ubiquinone is a redox-active prenol localized in the membranes [1] and conserved eukaryotes and many proteobacteria. Ubiquinone is composed of two main parts, a redox-active aromatic group forming a polar head and a hydrophobic polyisoprenoid tail [1]. In *Escherichia coli,* the biosynthesis of the ubiquinone is performed via a cytosolic complex, called the Ubi metabolon [2]. This latter is composed of different subunits including enzymes (methyltransferases UbiG and E, hydroxylases UbiI, H and F) and structural proteins (UbiJ and K). UbiJ is assumed to bind the hydrophobic tail via a domain called SCP2 located at the N-term, a domain known to interact with the lipid bilayer, notably to enable lipid transport [3]. Moreover, experimental evidences have revealed an interaction between UbiJ and an UbiK dimer, but also between UbiK and the cellular membrane in *E. coli* [2]. Considering the biological context and the structural importance of UbiJ and UbiK, we investigated the molecular complex UbiJ-UbiK$_2$ using multiscale molecular modelling approaches.

Our *in silico* study is articulated around three main aspects: (i) modelling of the protein partners (UbiJ and UbiK), (ii) identifying interactions of the complex with the membrane via molecular dynamics simulations, (iii) investigating binding of ubiquinone inside the UbiJ-UbiK$_2$ complex and its release in the membrane. Due to the lack of 3D template, AlphaFold2 [4] was used to predict individually the 3D models of UbiJ, UbiK, and their assembly which were further assessed using the contacts predicted from coevolution information. A multiscale approach combining both Coarse-Grained (Martini3 Force Field [5]) and all-atom (CHARMM36 [6]) modelling methods was used to identify the molecular interactions of the UbiJ-UbiK$_2$ complex with the membrane. Finally, the release of the ubiquinone was quantitatively estimated using biased molecular dynamics simulations such as Umbrella Sampling.

The main results of this work are: (i) the identification of the key amino acid residues involved in the interaction between UbiJ-UbiK$_2$ complex and the membrane, validated by different scales of molecular dynamics simulations, (ii) the validation of the ubiquinone binding mode within the UbiJ-UbiK$_2$ complex through free energy calculations along its release towards the membrane.

Acknowledgements

The PhD of R. Launay is funded by a grant from the Ministry of Education, Research and Innovation (MESRI). This work was performed using HPC resources from CALMIP (Grant 2022-p19013).

References

1. J.A Stefely and D. J. Pagliarini. Biochemistry of Mitochondrial Coenzyme Q Biosynthesis. *Trends in Biochemical Sciences*, (42): 824-843, 2017.
2. M. Hajj Chehade et al. A soluble metabolon synthesizes the isoprenoid lipid ubiquinone. *Cell Chemical Biology*, (26): 482-492, 2019.
3. N. Burgardt et al. A structural appraisal of sterol carrier protein 2. *Biochimica et Biophysica Acta*, (1865): 565-577, 2017.
4. J. Jumper et al. Highly accurate protein structure prediction with AlphaFold. *Nature* (596): 583-589, 2021.
5. P.C.T Souza et al. Martini3: a general purpose force field for coarse-grained molecular dynamics. *Nat Methods* (18): 382-388, 2021
6. J. Huang and A.D. MacKerell Jr. CHARMM36 all-atom additive protein force field: Validation based on comparison to NMR data. *Journal of Computational Chemistry* (34): 2135-2145, 2013.

Getting two birds with one stone: The Bios2cor R package for protein correlation analysis

Bruck TADDESE[1], Antoine GARNIER[1], Madeline DENIAUD[1], Daniel HENRION[1] and Marie CHABBERT[1]

[1] MITOVASC Laboratory, UMR CNRS 6015 – INSERM U1083, Université d'Angers, 49100 ANGERS, FRANCE

Corresponding Author: marie.chabbert@univ-angers.fr

1 Introduction

The analysis of covariation in multiple sequence alignments (MSA) of evolutionary related proteins is a widely used method to gain information on protein structure and function. The analysis of correlated sidechain motions in protein molecular dynamics simulations (MD) might also provide valuable information on the structural dynamics underlying protein functions. Here, using the Bios2cor R package that we have developed, we show that integration of both types of data yields identification of functionally important residues in the G protein-coupled receptor family.

2 Method

The R package Bios2cor (https://CRAN.R-project.org/package=Bios2cor) has been developed for the analysis of covariation/correlation data from both multiple sequence alignments and MD simulations [1]. Using a variety of scoring functions, Bios2cor computes covariation/correlation scores between either positions in an MSA or dihedral angles/rotamers in MD simulations. In addition, it provides friendly tools for data representation and interpretation, including network and principal component analyses.

3 Results

We have applied the Bios2cor package to the G protein-coupled receptor family, to gain insights into the mechanism of action of chemokine receptors. We analyzed sequence covariation in different hierarchical sequence sets ranging from the GPCR family to the chemokine receptor sub-family. This approach revealed the key role of position 3.35, located at the allosteric sodium binding site, in the functional evolution of chemokine receptors [2]. Then we carried out accelerated molecular dynamics simulations of the chemokine receptor CXCR4. During these simulations, we observed a conformational transition of CXCR4 from an inactive to an active-like conformation. Analysis of correlated sidechain motions during this transition identified a set of residues whose collaborative sidechain rotamerization immediately preceded or accompanied the conformational transition. In particular, the reorientation of the residue at position 3.35 should be crucial for receptor activation [3].

4 Conclusion

Bios2cor is a user-friendly package aimed at analyzing and interpreting covariation/correlation data from both protein sequences and MD simulations. The chemokine receptor family provides an example of an evolutionary important residue whose conformational reorientation plays a pivotal role in protein function. Bios2cor integrates both evolutionary and dynamical data to highlight key residues in the functional evolution of protein families.

Acknowledgements

This study was supported by the ANR Chemo_Tx_ProG and GENCI 100567 grants to MC.

References

[1] Bios2cor: an R package integrating dynamic and evolutionary correlations to identify functionally important residues in proteins. Taddese B, Garnier A, Deniaud M, Henrion D, Chabbert M. (2021) Bioinformatics: btab002.

[2] Evolution of chemokine receptors is driven by mutations in the sodium binding site. Taddese B, Deniaud M, Garnier A, Tiss A, Guissouma H, Abdi H, Henrion D, Chabbert M. (2018) PLoS Comput Biol 14(6):e1006209.

[3] Deciphering collaborative sidechain motions in proteins during molecular dynamics simulations. Taddese B, Garnier A, Abdi H, Henrion D, Chabbert M. (2020) Sci Rep. 10(1):15901.

A new way for finding drugs target protein and discover new protein complex in CETSA experiment

Marc-Antoine GERAULT[1,2], Pär Nordlund[3], Luc CAMOIN[1], Samuel GRANJEAUD[1]

[1] Aix-Marseille University, INSERM, CNRS, Institut Paoli-Calmettes, CRCM, Marseille Protéomique, F-13009, Marseille, France

[2] Centrale Marseille School, F-13013, Marseille, France

[3] Pär Nordlund's Group, Karolinska Institutet, S-17164, Stockholm, Sweden

Corresponding Author: marco.gerault@gmail.com

Background: The Cellular Thermal Shift Assay (CETSA) (originally described in [1]) is a biophysical assay based on the principle of ligand-induced thermal stabilization of target proteins, meaning that a protein's melting temperature will change upon ligand interaction. By heating samples (lysate, cells or tissue pieces) to different temperatures (typically 10) and quantifying proteins expression, we can detect altered protein interactions after for example drug treatment. The thermal profile of each protein is established by fitting a sigmoid curve on these ten points, which allows determining its melting temperature. Such an experiment is complex and time consuming to set up, but above all hard to analyze. A simpler way described in [2] is to consider only the expression fold change between the treatment and the control. Thus, only six temperatures (or even less) are needed and the sigmoid fit is unnecessary. Although the proof of concept was published in [2], the numerical analysis still needs to be validated and a robust scoring has to be determined in order to focus the analysis on proteins with the largest change in melting temperature. In addition, it would be interesting to identify proteins with similar CETSA profiles because they may belong to the same protein complex. Finally, to make these analyses easily accessible to the experimenter, an interactive graphical exploration interface would be useful.

Methods: To determine a score for ranking the best hits, we first evaluated the mean of differences as in [3], although the mean can lower the score if a single temperature has a significant fold change. However, a hit can only have one significant fold change. Instead, we chose to compute a weighted least-squared regression on the absolute value of the fold changes, ranked in the decreasing order with larger weights on the two largest fold changes. From this regression, we obtain the intercept which will be the score, called *Stability Rate (SR)*. In this way, the score is always positive, which overcomes the problem explained earlier. We can now plot the *combined p-value* (Fisher's test p-value of the two p-values of the two biggest fold changes) against the *SR.* The more the protein is in the top right corner of the plot, the more confident we can be that the drug is binding to this protein. To find proteins with similar thermal profiles, we also chose to compute a *similarity score*. A simple score is the *Euclidean distance score*, which is between 0 and 1. The closer it is to one, the more similar the profiles are. By setting a cutoff, we can identify those proteins. To evaluate the relevance of their associations or their belonging to a complex, we can query the STRING database.

Results: We tested these two methods on the elutriation dataset from [2]. With a *combined p-value* cutoff of 0.01 and an *SR* cutoff of 0.5 we found 501 hits in total with ten times fewer hits in G2 phase compare to S phase (as in [2]). For example CCNB1 comes in first position followed by CDK1 as shown in [2]. In S phase it's RFC3. By searching profiles similar to RFC3 (*cutoff 0.3*) we recovered the RFC complex. For EXOSC4, the similarity search recovered only EXOSC1. With the vimentin protein (VIM) we recovered a STRING network of 6 interactions, a potential complex formed by VIM, VCP, ACTN1 and YWHAG. This technique could be a new way to focus faster on proteins revealed by CETSA and to identify protein complex. To ease the analysis of CETSA experiments in a single and user-friendly way, we developed an R package mineCETSAapp (https://github.com/marseille-proteomique/mineCETSAapp) that offers a Shiny application.

References

1. Daniel Martinez Mo et al. *Monitoring Drug Target Engagement in Cells and Tissues Using the Cellular Thermal Shift Assay.* Science, 341(6141):84-87, 2013.
2. Lingyun Dai et al. *Modulation of Protein-Interaction States through the Cell Cycle.* Cell 173, 1481–1494, 2018.
3. Martinez-Val et al. *Spatial-proteomics reveals phospho-signaling dynamics at subcellular resolution*. Nature Communications, 12:7113, 2021.

Molecular modeling of plasmodesm organization by MCTP proteins

Sujith Sritharan[1], Emmanuelle Bayer[2] and Antoine Taly[1]

[1] Laboratoire de Biochimie Théorique, UPR9080, CNRS, Université Paris Cité, Paris, France

[2] Laboratoire de Biogenèse Membranaire, UMR5200, CNRS, Université de Bordeaux, Villenave d'Ornon, France

Corresponding Author: sujith.sritharan@etu.u-paris.fr

Abstract

In plants, intercellular communication is primarily achieved through plasmodesmata. These membrane pores cross the cell wall and create symplastic continuity between cells [1]. Plasmodesms are crucial in coordinating developmental processes and defense mechanisms against pathogens.[2] They are also hijacked by viruses that can structurally modify them to propagate their viral genome from cell to cell.

Plasmodesms have a unique membrane organization: they are crossed by a "tube" of endoplasmic reticulum (ER), which is in intimate contact with the plasma membrane (PM), delimiting the pores. The two membranes are only a few nm apart (~10 nm) and connected by "tethers". The multiple C2 domains and transmembrane region protein (MCTP) family, critical regulators of cell-to-cell signaling in plants, act as ER-PM tethers, specifically at plasmodesmata [2,3]. However, the molecular mechanism and function of membrane tethering within plasmodesmata remain unknown. Furthermore, MCTP proteins are still poorly known at the level of the 3D structure.

Thus we used, AlphaFold, an innovative tool had been developed by DeepMind, which combines deep learning and graph theory that predicts the 3D structure of the protein [4]. We used the MCTP4 transmembrane protein model produced with alphafold2 as a starting point. We then run MD with a coarse-grained representation, using the MARTINI3 force-field, to provide a molecular description of MCTP interacting with ER-PM membranes.

References

[1] K. Knox *et al.*, « Putting the Squeeze on Plasmodesmata: A Role for Reticulons in Primary Plasmodesmata Formation », *Plant Physiol.*, vol. 168, nº 4, p. 1563‑1572, août 2015, doi: 10.1104/pp.15.00668.

[2] « Multiple C2 domains and transmembrane region proteins (MCTPs) tether membranes at plasmodesmata », *EMBO Rep.*, vol. 20, nº 8, p. e47182, août 2019, doi: 10.15252/embr.201847182.

[3] J. D. Petit, Z. P. Li, W. J. Nicolas, M. S. Grison, et E. M. Bayer, « Dare to change, the dynamics behind plasmodesmata-mediated cell-to-cell communication », *Curr. Opin. Plant Biol.*, vol. 53, p. 80‑89, févr. 2020, doi: 10.1016/j.pbi.2019.10.009.

[4] J. Jumper *et al.*, « Highly accurate protein structure prediction with AlphaFold », *Nature*, vol. 596, nº 7873, Art. nº 7873, août 2021, doi: 10.1038/s41586-021-03819-2.

Reduced structural flexibility of eplet amino acids in HLA proteins

Diego Amaya-Ramirez[1], Romain Lhotte[2] , Magali Devriese[2], Constantin Hays[2], Jean-Luc Taupin[2] and Marie-Dominique Devignes[1]

[1] Université de Lorraine, CNRS, Inria, LORIA, Nancy, F-54000, France
[2] INSERM, Hôpital Saint Louis, F-75010 Paris

Corresponding Author: diego.amaya-ramirez@inria.fr

Abstract :

The proteins encoded by the HLA (Human Leukocyte Antigen) system play a fundamental role in the immune response. However, in the context of organ transplantation, these proteins are the main reason for the loss of transplants, largely due to their omnipresence in nucleated cells and their great diversity in terms of sequence (the genes encoding these proteins are the most polymorphic genes in humans). Although experimental methods are now available to assess the presence of antibodies against donor HLA proteins in recipient's serum, using a panel of about 200 purified HLA proteins, the molecular mechanisms governing such recognition remain unknown.

A recent study [1] has demonstrated the importance of reduced structural flexibility to identify favorable regions in antigens for antibody binding. Amino acid flexibility is expressed through the Normalized Root Mean Square Fluctuation (N-RMSF) during a molecular dynamics simulation. However this study is restricted to 61 proteins where sequence dissimilarity was favored and where no HLA protein is found. In the present work we tested the underlying hypothesis in [1] applied to the particular case of the HLA system, which has the remarkable peculiarity of high sequence similarity. Moreover, we were interested in studying the difference in flexibility at the eplet rather than epitope level, that is to say at the level of the polymorphic amino acids that are characteristics of specific HLA proteins.

For this purpose, we performed molecular dynamics simulations of 203 HLA proteins. The starting structures for these simulations were either obtained from the PDB [2] (68 proteins) or modeled by Alphafold2 [3] (135 proteins). The N-RMSF of residues having at least a Relative Solvent Accessible Surface Area (RSASA) of 20% was then calculated. The list of confirmed eplets in the HLA system was obtained from EpRegistry [4] and any residue that does not appear as a possible eplet in EpRegistry was considered a non-eplet residue. The Kolmogorov-Smirnov test was used to verify whether there is a statistically significant difference between the two distributions of N-RMSF (eplet residues vs non-eplet residues).

As a result, we found that there is a significantly reduced flexibility for eplet residues compared with non-eplet residues. This result opens the door to the use of structural flexibility to identify antibody binding sites on HLA proteins.

Acknowledgements

Experiments presented in this paper were carried out using the MBI-DS4H and Grid-500 platforms. Grid'5000 testbed is supported by a scientific interest group hosted by Inria and including CNRS, RENATER and several Universities as well as other organizations (see https://www.grid5000.fr). Both platforms hardware has been funded in part by the CPER IT2MP (Contrat Plan État Région, Innovations Technologiques, Modélisation & Médecine Personnalisée) and FEDER (Fonds européen de développement régional). DAR is recipient of an Inria-Inserm PhD contract

References

1. D.G. Kim, Y. Choi, H.S. Kim, « Epitopes of Protein Binders Are Related to the Structural Flexibility of a Target Protein Surface », *J. Chem. Inf. Model.*, 2021
2. H. M. Berman *et al.*, « The Protein Data Bank », *Nucleic Acids Research*, vol. 28, nº 1, p. 235‑242, janv. 2000.
3. J. Jumper, R. Evans, *et al.*, « Highly accurate protein structure prediction with AlphaFold », *Nature*, 2021.
4. R.J. Duquesnoy, M. Marrari, L. Sousa et al., « Workshop report : a website for the antibody-defined HLA epitope registry », *International Journal of Immunogenetics*, 2012.

Investigating structural and sequence determinants of regioselectivity and substrate specificity in a family of enzymes: the case study of ubiquinone biosynthesis hydroxylases

Elin Teppa[1], Romain Launay[1], Alexandre G de Brevern[2], Ivan Junier[3], Sophie Abby[3], Fabien Pierrel[3], Jérémy Esque [1*], Isabelle André[1*]

[1] Toulouse Biotechnology Institute, TBI, Université de Toulouse, CNRS, INRAE, INSA, Toulouse, France. 135, avenue de Rangueil, F-31077 Toulouse Cedex 04, France
[2] Biologie Intégrée du Globule Rouge UMR_S 1134, Inserm, Laboratoire d'Excellence GR-Ex, Université de Paris, F-75739 Paris, France
[3] TIMC, Université Grenoble Alpes, CNRS, CHU Grenoble Alpes, Grenoble INP, TIMC-IMAG, Université Grenoble Alpes, Grenoble, France

Corresponding Author: isabelle.andre@insa-toulouse.fr, esque@insa-toulouse.fr

Here we present a computational approach applied to characterize the active site of a family of enzymes showing variation in their specificity and regioselectivity. We aim to understand the structural and molecular bases governing the differences in substrate specificity and the determinants of the narrow/broad regioselectivity in a family of enzymes involved in the biosynthesis of ubiquinone (UQ). We focused on four Ubi subfamilies responsible for the three hydroxylation reactions of UQ biosynthesis [1], which present both differences in substrate specificity and regioselectivity. All subfamilies belong to class A flavin-dependent mono-oxygenases (FMOs).

To understand the structural basis for this distinct regioselectivity, we analyzed UbiI [2], UbiL, and UbiM proteins showing a narrow, intermediate, and broad regioselectivity respectively [1]. It has been demonstrated that some organisms, such as *E. coli,* contain three UQ hydroxylases, each of one hydroxylates a specific position of the UQ aromatic ring (C5, C1 and C6). However, other organisms (i.e. *Rhodobacter capsulatus*) contain two enzymes to catalyze the three reactions: UbiL hydroxylates C-5 and C-1 showing an intermediate level of regioselectivity; whereas UbiN hydroxylates C-6 position. There are also organisms containing only UbiM protein to perform the three reactions. For instance, *Neisseria meningitidis* presents a single UbiM protein able to hydroxylate C-1, C-5, and C-6 positions, showing broad regioselectivity.

Our work is naturally split into two parts, the first is dedicated to substrate specificity and the second part to regioselectivity, (i) Identification of substrate specificity determining positions [3], (ii) Identification of residues involved in regioselectivity.

To conclude, our study provides important insights into the sequence-structure-function relationships of Ubi hydroxylase subfamilies.

Acknowledgements

This work was funded by the French National Research Agency (ANR Project Deepen, *ANR-19-CE45-0013-02*). This work was granted access to the HPC resources on the TGCC-Curie, CINES-Occigen supercomputers and the Computing mesocenter of Région Midi-Pyrénées (CALMIP, Toulouse, France).

References

1. L. Pelosi et al. Evolution of Ubiquinone Biosynthesis: Multiple Proteobacterial Enzymes with Various Regioselectivities To Catalyze Three Contiguous Aromatic Hydroxylation Reactions. *mSystems*, (4), 2016.
2. M. Hajj Chehade et al. ubiI, a New Gene in Escherichia coli Coenzyme Q Biosynthesis, Is Involved in Aerobic C5-Hydroxylation. *Journal of Biochemical Chemistry*, (27): 20085-20092, 2013.
3. A. Chakraborty et al. SPEER-SERVER: A Web Server for Prediction of Protein Specificity Determining Sites. *Nucleic Acids Research* 40 (Web Server issue): W242–48,2012.

BioacPepFinder: Discovery of bioactive peptides from protein digestion

Carlos BATHICH[1,4*], Isabelle GUIGON[2*], Hélène TOUZET[3], Rozenn RAVALLEC[1] and Christophe FLAHAUT[1]

[1] UMR Transfrontalière BioEcoAgro N° 1158, Univ. Lille, Univ. Artois, INRAE, Univ. Liège, UPJV, JUNIA, Univ. Littoral Côte d'Opale, ICV – Institut Charles Viollette, F-59000 Lille, France

[2] Univ. Lille, CNRS, Inserm, CHU Lille, Institut Pasteur de Lille, US 41-UAR 2014-PLBS-Plateforme bilille, F-59000 Lille, France

[3] Univ. Lille, CNRS, Centrale Lille, UMR 9189 CRIStAL, F-59000 Lille, France

4 Copalis Industrie, 62480 Le Portel, France

*Co-first authors. Corresponding Author: isabelle.guigon@univ-lille.fr

For several decade, health and alternative medicine sectors focus their research on the bioactive peptides (BAPs) derived from food, due to their preventive and/or health benefit applications. However, the experimental approaches for BAP discovery remain long and challenging. As a consequence, *in silico* methods proved to be more strategic due to the time and cost savings, and their efficient prediction of potential BAPs [1].

In this context, we developed a comprehensive pipeline, called BioacPepFinder, that is able to perform large-scale screening in order to discover new BAPs from *in silico* enzymatic hydrolysis of proteins. This pipeline takes as input a set of amino acid sequence of proteins (in FASTA format or with Ensembl, Uniprot or NCBI id). The first step is to simulate *in silico* hydrolysis, with RPG [2] coupled to an in-house script to deal with missed cleavages. The list of generated peptides is then filtered according to two complementary criteria. Peptides are compared with Blast to specialized databases of BAPs such as BIOPEP [3] and DRAMP [4], that contain known peptide sequences displaying proven bioactivities. We also compute for each peptide the quantitative structure-activity relationship (QSAR) score for discovering novel and potent BAPs [5]. This score is based on crucial structural properties, targets peptides with angiotensin-converting enzyme (ACE) and dipeptidyl peptidase-IV (DPP-IV) inhibitory activities.

BioacPepFinder is implemented in Python and dependencies are managed with Conda. It is modular, as each step can be carried out separately. It is also extensible: it is possible to add new proteases (for the hydrolysis), new databases or new criteria to select peptides (with wrappers). It is freely accessible at https://gitlab.univ-lille.fr/bilille/bioacpepfinder.

We have evaluated the BioacPepFinder capability of BAP prediction on standard proteins such as bovine-hemoglobin or -serum albumin using the BIOPEP database and QSAR dedicated to ACE and DPP-IV. It proved to be a new high-efficient prediction software that aims for guiding development and optimization of BAPs in order to advance in their production.

References

1. Udenigwe CC. Bioinformatics approaches, prospects and challenges of food bioactive peptide research. Trends Food Sci Technol. 2014;36(2):137–43. https://doi.org/10.1016/j.tifs.2014.02.004
2. Maillet N. (2019). Rapid Peptides Generator: fast and efficient in silico protein digestion. NAR genomics and bioinformatics, 2(1), lqz004. https://doi.org/10.1093/nargab/lqz004
3. Minkiewicz, P., Iwaniak, A., & Darewicz, M. (2019). BIOPEP-UWM Database of Bioactive Peptides: Current Opportunities. International journal of molecular sciences, 20(23), 5978. https://doi.org/10.3390/ijms20235978
4. Shi, G., Kang, X., Dong, F., Liu, Y., Zhu, N., Hu, Y., Xu, H., Lao, X., & Zheng, H. (2022). DRAMP 3.0: an enhanced comprehensive data repository of antimicrobial peptides. *Nucleic acids research*, *50*(D1), D488–D496. https://doi.org/10.1093/nar/gkab651
5. Nongonierma, A. B.; FitzGerald, R. J. Learnings from quantitative structure-activity relationship (QSAR) studies with respect to food protein-derived bioactive peptides: a review. RSC Adv. 2016, 6, 75400−75413.

Towards the potentiation of selective inhibition of Mfd nanomachine

Samantha Samson[1], Delphine Cormontagne[2], Seav-Ly Tran[2], Lucie Lebreuilly[2], Jean-Christophe Cintrat[3], Didier Rognan[4], Nalini Rama Rao[2] and Gwenaëlle André[1]

[1] MaIAGE, INRAE, Université Paris-Saclay, 78350, Jouy-en-Josas, France
[2] Micalis, INRAE, AgroParisTech, Université Paris-Saclay, 78350, Jouy-en-Josas, France
[3] LCB, CEA, Université Paris-Saclay, 91191, Gif-sur-Yvette, France
[4] LIT, CNRS, Université de Strasbourg, 67400, Illkirch, France

Corresponding Author: samantha.samson@inrae.fr

Mutation frequency decline (Mfd) is a 120 kDa multifunctional protein, ubiquitous in all bacteria. It has been recently shown to boost antibiotic resistance, as it fosters the appearance of bacterial DNA mutations [1]. Additionally, it is a virulence factor that is overexpressed in bacterial cells to overcome DNA damages, after oxidative stress such as nitric oxide (NO), produced by macrophages. As such, it acts as a key player to cancel out the innate immune response [2]. Through cycles of ATP hydrolysis, Mfd translocates onto the template DNA strand and promotes its disassembly from the RNAP [3]. Hence, we focused on Mfd ATP-ase function to identify molecules capable of competing with ATP and blocking this step.

Virtual screening of 5 million compounds was computed on RecG, a homologous protein, that was the only protein whose coordinates were available in an active conformation at that time [4]. A follow-up through a medium-throughput activity test performed on Mfd from *Escherichia coli* confirmed compounds as very effective *in vitro*. A first selection of hits was refined through docking, using AutoDock Tools [5], on Mfd from *E. coli* whose coordinates were solved [6]. Then, this work was extended to the Mfd of the ESKAPE group, considered by the World Health Organization, as priority pathogens. Interestingly, those hits display a much better affinity as compared to ATP/ADP and conserved positions were identified as key for the affinity binding. Currently, *in vivo* validation is in progress. Markedly, 6 promising compounds have been identified, one of which was already tested *in vivo* and effectively treats Gram-negative infection in animal models. Here, we propose to describe our work and discuss our promising results.

References

[1] M. N. Ragheb et al., « Inhibiting the Evolution of Antibiotic Resistance », Mol. Cell, vol. 73, no 1, p. 157-165.e5, janv. 2019.
[2] E. Guillemet et al., « The bacterial DNA repair protein Mfd confers resistance to the host nitrogen immune response », Sci. Rep., vol. 6, no 1, Art. no 1, juill. 2016.
[3] J. Y. Kang et al., « Structural basis for transcription complex disruption by the Mfd translocase », eLife, vol. 10, p. e62117, janv. 2021.
[4] M. R. Singleton, S. Scaife, et D. B. Wigley, « Structural analysis of DNA replication fork reversal by RecG », Cell, vol. 107, no 1, p. 79-89, oct. 2001.
[5] G. M. Morris et al., « AutoDock4 and AutoDockTools4: Automated Docking with Selective Receptor Flexibility », J. Comput. Chem., vol. 30, no 16, p. 2785-2791, déc. 2009.
[6] A. M. Deaconescu et al., « Structural basis for bacterial transcription-coupled DNA repair », Cell, vol. 124, no 3, p. 507-520, févr. 2006.

Structural prediction of macromolecular interactions using evolutionary information

Chloé QUIGNOT[1], Hélène BRET[1], Ikram MAHMOUDI[1], Raphael GUEROIS[1] and Jessica ANDREANI[1]

[1] Université Paris-Saclay, CEA, CNRS, Institute for Integrative Biology of the Cell (I2BC), 91198, Gif-sur-Yvette, France

Corresponding Author: jessica.andreani@cea.fr

1. Introduction

The vast majority of biological processes rely on macromolecular interactions. Protein docking aims to predict the most likely structural binding modes of interacting partners. Using information about the coevolution of protein partners improves the success rate of these predictions. Our team has contributed to the improvement of protein docking strategies through such incorporation of evolutionary information. Here, we present our recent achievements in this respect, as well as ongoing work and perspectives.

2. The InterEvDock3 docking pipeline

In previous work, we had developed a protein docking pipeline that integrates evolutionary information [1]. Recently, we designed a novel strategy to integrate evolutionary information into atomic-level scoring methods, using shallow multiple sequence alignments, and we found that it greatly improved their capacity to discriminate correct from incorrect interface models [2]. We integrated this strategy into the InterEvDock3 docking server [3], along with the capacity to use covariation-based contact maps (possibly derived from deep-learning-based strategies) and the ability to combine free docking with template-based assembly modeling. We also successfully applied the InterEvDock pipeline to recent targets of the CAPRI assembly prediction challenge [4]. The server is available at https://bioserv.rpbs.univ-paris-diderot.fr/services/InterEvDock3/

3. Ongoing work

Ongoing work in our team aims at developing a deep learning method for the protein docking scoring step by integrating the physical-chemical environment of amino acids, the interface geometry, and sequence or evolutionary information if available. Another project aims at extending the use of evolutionary information to the structural prediction of protein-RNA interactions, as well as benefiting from high-throughput protein-RNA "interactomics" data to enrich available high-resolution structural data.

Acknowledgements

This work was supported by the French National Research Agency under grants ANR-15-CE11-0008 to R.G. and ANR-18-CE45-0005 to J.A., by CEA doctoral funding to H.B. and by IDEX Paris-Saclay doctoral funding to C.Q. It was granted access to the HPC resources of CCRT under allocations 2018-7078 and 2019-7078 by GENCI (Grand Equipement National de Calcul Intensif).

References

1. Chloé Quignot, Julien Rey, Jinchao Yu, Pierre Tufféry, Raphael Guerois and Jessica Andreani. InterEvDock2: an expanded server for protein docking using evolutionary and biological information from homology models and multimeric inputs. *Nucleic Acids Res*, 46(W1):W408-W416, 2018.
2. Chloé Quignot, Pierre Granger, Pablo Chacón, Raphael Guerois and Jessica Andreani. Atomic-level evolutionary information improves protein-protein interface scoring. *Bioinformatics*, 37(19): 3175–3181, 2021.
3. Chloé Quignot, Guillaume Postic, Hélène Bret, Julien Rey, Pierre Granger, Samuel Murail, Pablo Chacón, Jessica Andreani, Pierre Tufféry and Raphaël Guerois. InterEvDock3: a combined template-based and free docking server with increased performance through explicit modeling of complex homologs and integration of covariation-based contact maps. *Nucleic Acids Research*, 49(W1):W277–W284, 2021.
4. Aravindan Arun Nadaradjane, Chloé Quignot, Seydou Traoré, Jessica Andreani and Raphael Guerois. Docking proteins and peptides under evolutionary constraints in Critical Assessment of PRediction of Interactions rounds 38 to 45. *Proteins*, 88(8):986-998, 2020.

Modeling of NK-cell immunosurveillance in normal and pathological BM microenvironnement

Berenice Schell[1], Valeria Bisio[1], Lin Pierre Zhao[1], Emilie Lereclus[1], Camille Kergaravat[1], Emmanuel Clave[1], Antoine Toubert[1], Pierre Fenaux[1], Marion Espéli[1], Karl Balabanian[1], Lionel Adès[2] and Nicolas Dulphy[1]

[1] INSERM UMRS1160, Hôpital Saint-Louis AP-HP, Paris, France

[2] Hématologie sénior, Hôpital Saint-Louis AP-HP, Paris, France

Corresponding author: berenice.schell@inserm.fr

1 Introduction

Myelodysplastic syndromes (MDS) are clonal disorders of the Bone Marrow (BM) that evolve in secondary Acute Myeloid Leukemia (sAML), characterized by an immune escape and leukemic cell invasion of the BM, in 30% of cases [1]. In this context, the immunomodulatory Mesenchymal Stromal Cells (MSC) and the Natural Killer (NK) lymphocytes, both impaired in these pathologies, are known to play an active part in the disease evolution [2][3]. In this study, we wondered to what extent the defects observed in the NK-cell immunosurveillance are attributable to MSC immunosupressive role.

2 Material and Methods

To better reproduce cell-cell and cell-matrix interactions in the BM niche, we developed a 3D *in vitro* cell culture system putting together NK-cells, MSC and leukemic cells. NK-cells were isolated from healthy donor (HD) peripheral blood mononucleated cells (n=3). MSC were expanded *ex vivo* from the BM samples of sAML patients or HD that underwent a hip replacement (n=3). An increasing number of Molm13, a sAML cell line, has been added in order to modelize the progressive leukemic cell invasion. After two days of co-culture, the organoid was dissociated and the three cell types were characterized by spectral cytometry using Aurora, CYTEK®. After a preliminary gating using FlowJO 10.8.1, FCS files were processed through a pipeline developed in R 4.1.2 using RStudio comprising: arcsin transformation using FlowVS package, batch effect correction using a FlowSOM based algorithm, normalization [4] and clustering for trajectory inference analysis [5].

3 Results

This preliminary data allowed us to observe a down-regulation of NK-cell activating receptors when they are cultured with sAML compared with HD-MSC. In parallel, differential activation trajectory of NK-cells depending on the BM microenvironement was assessed. This allowed us to reconstruct a three cell interactome *in silico* and observe the NK-cell immune defects induced by the MSC that can be responsible for leukemic cell proliferation, BM invasion and finally disease progression.

References

[1] Lionel Adès, Raphael Itzykson, and Pierre Fenaux. Myelodysplastic syndromes. *The Lancet*, 383(9936):2239–2252, June 2014.

[2] Dhifaf Sarhan, Jinhua Wang, Upasana Sunil Arvindam, Caroline Hallstrom, Michael R. Verneris, Bartosz Grzywacz, Erica Warlick, Bruce R. Blazar, and Jeffrey S. Miller. Mesenchymal stromal cells shape the MDS microenvironment by inducing suppressive monocytes that dampen NK cell function. *JCI insight*, 5(5), March 2020.

[3] Pearlie K. Epling-Burnette, Fanqi Bai, Jeffrey S. Painter, Dana E. Rollison, Helmut R. Salih, Matthias Krusch, Jianxiang Zou, Edna Ku, Bin Zhong, David Boulware, Lynn Moscinski, Sheng Wei, Julie Y. Djeu, and Alan F. List. Reduced natural killer (NK) function associated with high-risk myelodysplastic syndrome (MDS) and reduced expression of activating NK receptors. *Blood*, 109(11):4816–4824, June 2007.

[4] Janine E. Melsen, Monique M. van Ostaijen-ten Dam, Arjan C. Lankester, Marco W. Schilham, and Erik B. van den Akker. A Comprehensive Workflow for Applying Single-Cell Clustering and Pseudotime Analysis to Flow Cytometry Data. *The Journal of Immunology*, 205(3):864–871, August 2020.

[5] Yuting Dai, Aining Xu, Jianfeng Li, Liang Wu, Shanhe Yu, Jun Chen, Weili Zhao, Xiao-Jian Sun, and Jinyan Huang. CytoTree: An R/Bioconductor package for analysis and visualization of flow and mass cytometry data. *BMC Bioinformatics*, 22(1):138, March 2021.

Investigation of neural population-based optimization

Vaitea Opuu[1]
MPI for mathematics in the Sciences, 04103, Leipzig, Germany

Corresponding author: vopuu@mis.mpg.de

We investigated in this work ways to accommodate deep learning techniques into evolutionary algorithms (EA) in order to reduce the level of expert tuning required to design such a method.

Global optimization is at the heart of countless applications in bioinformatics, where EA are among the method of choice. It has been applied to protein structure prediction, RNA sequence design, or molecular docking. Although it has shown success in most problems, the design of an efficient EA is challenging as it requires expert tuning, which is usually not transferable.

For the simplest EA, one starts with a random population of solutions for the problem, i) one applies a perturbation to the population, then ii) poor solutions are removed while good solutions are kept. Vanilla EAs use simple isotropic perturbations where each solution in the population is changed only a little in all "directions", which is usually called the mutational operator.

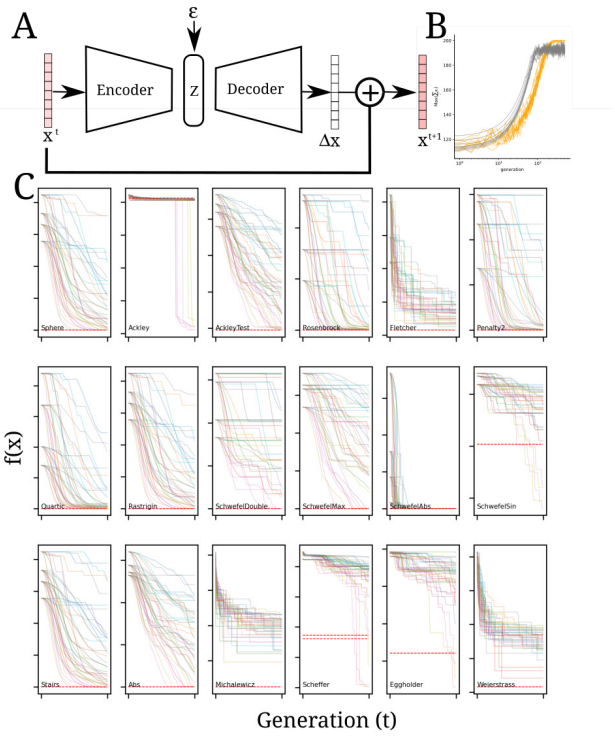


Fig. 1. Neural population-based optimization. A) description of the mutational operator where a solution x is encoded into a z vector using the network called "Encoder". A Gaussian noise ϵ is added to the vector z. Next, the decoder network is used to predict the perturbation to add to the solution. B) Comparison of a simple genetic algorithm (grey) and our method (in orange). Our method convergence is slower. C) Shows the performances of our method compared to differential evolution solution (red dotted line) for various network configurations (number of hidden units and layers, colored lines). For each function f, Y-axes show the minimum value $f(x)$ found in the population at each generation (x-axes).

Here, we investigated the accommodation of one type of neural network as a mutational operator. Starting with an un-trained network, we propose to learn optimal perturbations along the optimization process. To do so, we combined two types of architectures: 1) the auto-encoder that compresses the solutions into a lower-dimensional manifold and then decodes them, and 2) a recurrent network that is used to model time series. Similar approaches have been applied to enhance Monte Carlo simulation.

On a set of 18 well-known test functions, the proposed approach showed similar performances to the differential evolution algorithm implemented in scipy; however, using an easy-to-implement and straightforward concept. Next, we adapted our approach to discrete optimization. It showed encouraging performances on the one-max problem. However, the performances are mitigated for the more complex tasks such as RNA design or Ising model optimizations.

An implementation of our method is given at https://github.com/vaiteaopuu/neural_ga

Integrated Analyses of Large Scale RNAseq Data in Acute Myeloid Leukemia

Raíssa Silva[1,3], Cédric Riedel[1,2,3], Benoit Guibert[1], Anthony Boureux[1], Florence Ruffle[1] and Thérèse Commes[1]

[1] IRMB, University of Montpellier, INSERM U1183, 80 rue Augustin Fliche, 34295 Montpellier, France

[2] Faculty of Medicine, University of Montpellier, Montpellier, France

[3] Equal authors

Corresponding author: therese.commes@inserm.fr

Acute Myeloid Leukemia (AML) is a hematopoietic disorder characterized by the infiltration of the bone marrow, blood, and other tissues by clonal and poorly differentiated cells [1]. The AML is classified into sub-types that help to improve clinical outcomes and rate survival. Even though the current classifications are well-known and highly adopted such as French-American-British (FAB) [2] and European LeukemiaNet (ELN) [3], the misunderstanding to classify specific patients shows the need of new biomarkers for more confident classification and propose personalized treatment.

A large exploration of RNAseq data provides a deep investigation to identify and target key molecular mechanisms that drive the AML pathogenesis and progression. In this context, we developed a pipeline to explore the expression of the different genes using k-mer-based-approach and Machine Learning methods to yield a better understanding of prognosis classification. As initial tests, we used NPM1 mutation gene to search for linked genes that can show a new perspective of the disease. The pipeline was applied to differentiate the condition of mutated and non-mutated NPM1 patients. We used Kmtricks tool [4] to count the k-mers present in each RNAseq sample. Then, we used a feature selection step considering the coefficient of variation between the conditions, thus, allowing us to select relevant k-mers. After, we use the Machine Learning method to classify the conditions and find the more important k-mers based on the classification. Finally, we identified the genes corresponding to these k-mers with BLAT [5].

In our initial results, we found a differential expression of HOX family genes, mainly for HOXB9 gene. This gene showed high expression for mutated NPM1 patients and low expression for non-mutated NPM1 patients, in 5 AML different cohorts (744 samples). When analyzed in 132 wild-type samples, the expression was minimal. Furthermore, HOX genes are already known genes related to AML, giving us confidence that our results were not random and guiding us to apply the pipeline to new genes and prognosis.

Acknowledgements

This work was supported by La Ligue Contre le Cancer - France, and “Bourse Année de recherche médecine de la faculté de Médecine Montpellier-Nîmes, Université de Montpellier”.

References

[1] Hartmut Döhner, Daniel J Weisdorf, and Clara D Bloomfield. Acute myeloid leukemia. *New England Journal of Medicine*, 373(12):1136–1152, 2015.

[2] John M Bennett, Daniel Catovsky, Marie-Theregse Daniel, George Flandrin, David AG Galton, Harvey R Gralnick, and Claude Sultan. Proposals for the classification of the acute leukaemias french-american-british (fab) co-operative group. *British journal of haematology*, 33(4):451–458, 1976.

[3] Tobias Herold, Maja Rothenberg-Thurley, Victoria V Grunwald, Hanna Janke, Dennis Goerlich, Maria C Sauerland, Nikola P Konstandin, Annika Dufour, Stephanie Schneider, Michaela Neusser, et al. Validation and refinement of the revised 2017 european leukemianet genetic risk stratification of acute myeloid leukemia. *Leukemia*, 34(12):3161–3172, 2020.

[4] Téo Lemane, Paul Medvedev, Rayan Chikhi, and Pierre Peterlongo. kmtricks: Efficient construction of bloom filters for large sequencing data collections. 2021.

[5] W James Kent. Blat—the blast-like alignment tool. *Genome research*, 12(4):656–664, 2002.

A novel method to identify and score clusters of motifs of protein sequences (CLUMPs) based on amino acids physicochemical properties.

Paola Porracciolo[1,2*], Djampa Kozlowski[1,2*], Etienne Danchin[2#] and Silvia Bottini[1#]

[1] Université Côte d'Azur, Center of Modeling, Simulation and Interactions, Nice, France

[2] INRAE, Université Côte d'Azur, CNRS, Institut Sophia Agrobiotech, Sophia-Antipolis, France

*contributed as co-first author
#contributed as co-last author
Corresponding Author: silvia.bottini@univ-cotedazur.fr

Nowadays, motif discovery is routinely used in high-throughput studies, including at the protein level. Motif-calling algorithms can be divided into two categories: generative models and discriminant methods. Unlike generative models, discriminative methods compare two datasets to identify motifs that are present at a high rate in the positive set compared to the negative one. Although several efficient discriminative motif-finding methods have been developed so far [1], they have led to new challenges. The first issue concerns the degeneration of the motifs because they identify a high number of motifs having either high sequence redundancy or different sequence composition. Furthermore, different amino acids (AAs) can have similar physicochemical properties, thus different motif sequences can share similar properties. The second issue concerns the lack of a scoring function that considers the characteristics of the AAs composing the motifs rather than the occurrence rate only.

To address these challenges, we have developed a novel tool that identifies clusters of motifs of protein sequences (CLUMPs) and associates a score to each CLUMPs. This score encompasses the physicochemical properties of AAs and the motif occurrences. Briefly, it takes as input the list of discriminant motifs identified in a positive set compared to a negative set and performs motif clustering based on 16 features describing the physicochemical properties of AAs. Finally, CLUMPs are sorted by a novel score that considers these properties according to the ones of the protein sequences composing the positive dataset with respect to the negative set and a Jaccard index-based indicator.

We used our method to identify discriminant CLUMPs of effector proteins in nematodes. Plant-parasitic nematodes secrete effector proteins inside their plant hosts to manipulate their development, defense systems, metabolism and physiology [2]. However, accurate detection of effector proteins in nematode genomes is challenging. We focused on *Meloidogyne incognita* species because their effectors are the most characterized in the literature among plant-parasitic nematodes and due to their agroeconomic importance. We applied our tool to a positive set and a negative set composed respectively of 161 and 495 protein sequences. We identified 6 CLUMPs which occur at specific positions in the positive sequences but not in the negative ones. We showed that the first 3 CLUMPs co-occur significantly in the positive sequences, while no co-occurrences are observed in the negative set. To test the validity of our findings, we searched for the 6 CLUMPs in an extended set of 14 parasitic nematode species comprising 624 and 4214 proteins in the positive and negative sets, respectively. Overall, we achieved similar results, allowing to conclude that our method identified discriminant CLUMPs for nematode effector proteins.

References

1. He Y, *et al.* A survey on deep learning in DNA/RNA motif mining. *Briefings in Bioinformatics* 22(4)bbaa229, 2021.
2. Vieira, *et al*. Plant-Parasitic Nematode Effectors — Insights into Their Diversity and New Tools for Their Identification. *Biotic Interactions* 50:37–43, 2019.

Prediction of tumor microenvironment heterogeneity in kidney cancers by cell deconvolution

Pauline BAZELLE[1], Sarah SCHOCH[2], Florian JEANNERET[1], HÅKAN AXELSON[2], Christophe BATTAIL[1], and KATY CONSORTIUM[3]

[1] Laboratoire Biologie et Biotechnologies pour la Santé, IRIG, UMR 1292 INSERM-CEA-UGA, Univ. Grenoble Alpes, 38000 Grenoble, France.

[2] Department of Laboratory Medicine, Lund University, Scheelevägen 2, 223 81 Lund, Sweden

[3] https://katy-project.eu, European Unions's Horizon 2020 research and innovation programme, Grant agreement No 101017453

Corresponding Author: pauline.bazelle@cea.fr

In the field of precision oncology, we seek, among other things, to determine the most appropriate drug therapy for each patient, taking advantage of the molecular and cellular profile of the tumor before treatment. This information makes it possible in particular to characterize the content of the tumor microenvironment, identified in recent years as being able to greatly influence patients response to targeted therapies, such as protein kinase inhibitors and immunotherapies.

The reference technology for measuring the cellular heterogeneity of the tumor microenvironment is single-cell RNA-seq. However, its experimental conditions of use, as well as its high financial cost, make it difficult to use it to profile large cohorts of patients. Indeed, they are rather analyzed with sequencing technologies such as bulk RNA-seq.

Cell deconvolution methods have emerged in recent years as relevant alternatives for predicting the proportions of cell types present in biological samples profiled by bulk RNA-seq. Within the framework of the European KATY project (https://katy-project.eu/), we are interested in the heterogeneity of the tumor microenvironment of the clear cell renal cell carcinoma (ccRCC) and its influence on the ability to predict a patient's response to a treatment. To meet this objective, we have optimized a bioinformatics protocol for the analysis of single-cell RNA-seq data and the prediction of cell fractions by deconvolution methods. The cell deconvolution methods used are CIBERSORTx [1] and MuSiC [2]. The single cell RNA-seq matrix [3] used to perform cell deconvolution was derived from 11 adult patients with ccRCC. We performed deconvolution on a bulk RNA-seq consisting of 311 ccRCC tumor samples [4]. We estimated the performance of our predictions using simulated data and by comparison with tumor purity scores.

References

[1] Newman, A. M., Steen, C. B., Liu, C. L., Gentles, A. J., Chaudhuri, A. A., Scherer, F., Khodadoust, M. S., Esfahani, M. S., Luca, B. A., Steiner, D., Diehn, M., & Alizadeh, A. A. (2019). Determining cell type abundance and expression from bulk tissues with digital cytometry. Nature biotechnology, 37(7), 773–782. https://doi.org/10.1038/s41587-019-0114-2

[2] Wang, X., Park, J., Susztak, K., Zhang, N. R., & Li, M. (2019). Bulk tissue cell type deconvolution with multi-subject single-cell expression reference. Nature communications, 10(1), 380. https://doi.org/10.1038/s41467-018-08023-x

[3] Obradovic A, Chowdhury N, Haake SM, et al. Single-cell protein activity analysis identifies recurrence-associated renal tumor macrophages. Cell. 2021;184(11):2988-3005.e16. doi:10.1016/j.cell.2021.04.038

[4] Braun, D. A., Hou, Y., Bakouny, Z., Ficial, M., Sant' Angelo, M., Forman, J., Ross-Macdonald, P., Berger, A. C., Jegede, O. A., Elagina, L., Steinharter, J., Sun, M., Wind-Rotolo, M., Pignon, J. C., Cherniack, A. D., Lichtenstein, L., Neuberg, D., Catalano, P., Freeman, G. J., Sharpe, A. H., … Choueiri, T. K. (2020). Interplay of somatic alterations and immune infiltration modulates response to PD-1 blockade in advanced clear cell renal cell carcinoma. Nature medicine, 26(6), 909–918. https://doi.org/10.1038/s41591-020-0839-y

Statistical approach for the detection of transposable element insertion bias in *Drosophila melanogaster*

Elyes BRAHAM*[1], Kateryna D.MAKOVA[2,3] and Anna-Sophie FISTON-LAVIER[1,4]
[1] ISEM, Univ Montpellier, CNRS, IRD, Montpellier, France
[2] Center for Medical Genomics, The Huck Institutes of the Life Sciences, Pennsylvania, United States
[3] Department of Biology, The Pennsylvania State University, University Park, USA.
[4] Institut Universitaire de France (IUF)

Corresponding author: elyes.braham@etu.umontpellier.fr*

Transposable elements (TEs) are mobile genetic elements that have the ability to duplicate, to transpose from one locus to another. They are ubiquitous sequences that can represent up to 90% of some genomes. TEs induce chromosomal rearrangements and genetic variability. By consequence, they can be involved in genome evolution such as adaptation processes. It is therefore important to study their dynamics to estimate their impact on such genomics processes. TE distribution along the genome varies depending on the balance between integration preference and post-insertion selection. Some studies highlighted TE insertion bias. For instance, in human, an insertion bias has been detected for L1 elements. Previous study suggests a preferentially insertion into open-chromatin regions enriched in non-B DNA motifs for those elements [1]. In order to analysis and characterize this insertion bias, Cremona et al further designed a statistical approach called IWTomics for *Interval-Wise-Testing for omics* (http://bioconductor.org/packages/IWTomics). This approach takes advantage of the numerous genomics data available to date (*e.g.* recombination rate, GC-rich, methylation) to statistically detect significant effects of the insertions along a genome. Genomics features can be treated as "curves" (mathematical functions) and then analyzed using functional data analysis (FDA) [2]. Using IWTomics, Chen et al presented the first high-resolution L1 transposition dynamics. They compared the TE landscapes for three different TE datasets : newly inserted elements, "*de novo*"; TEs inserted in some individuals of the populations, "polymorphic" , and genome specific TEs present in all individuals of the populations, "fixed". They confirm the L1 insertion bias and highlighted preference sites for *de novo* and fixed L1 insertions [3]. In *Drosophila melanogaster*, Merenciano et al identified a preferential insertion locus for *roo* elements within a promoter region of a stress response associated gene [4]. Such retrotransposon elements are known to be the most abundant and active ones in *D. melanogaster* [5]. Considering that a lot of genomics features are now available for this organism (https://flybase.org/), we decided to investigate the TE dynamics using IWTomics to test this approach. We started our study by focusing on roo elements and used three TE datasets ("*de novo*", "polymorphic" and "fixed" TEs), "*Drosophila*-specific elements". The aim of this exploratory work is to confirm if with IWTomics we can detect the already known insertion bias of roo elements in *D. melanogaster*. Next, we plan to develop a pipeline to automatize same analysis for all TE family to detect TE families with preferential insertion locus.

References

[1] Tania Sultana, Dominic van Essen, Oliver Siol, Marc Bailly-Bechet, Claude Philippe, Amal Zine El Aabidine, Léo Pioger, Pilvi Nigumann, Simona Saccani, Jean-Christophe Andrau, Nicolas Gilbert, and Gael Cristofari. The landscape of l1 retrotransposons in the human genome is shaped by pre-insertion sequence biases and post-insertion selection. *Molecular Cell*, 74(3):555–570.e7, 2019.

[2] Marzia A Cremona, Alessia Pini, Fabio Cumbo, Kateryna D Makova, Francesca Chiaromonte, and Simone Vantini. IWTomics: testing high-resolution sequence-based 'Omics' data at multiple locations and scales. *Bioinformatics*, 34(13):2289–2291, 02 2018.

[3] Di Chen, Marzia A Cremona, Zongtai Qi, Robi D Mitra, Francesca Chiaromonte, and Kateryna D Makova. Human L1 Transposition Dynamics Unraveled with Functional Data Analysis. *Molecular Biology and Evolution*, 37(12):3576–3600, 07 2020.

[4] Miriam Merenciano, Camillo Iacometti, and Josefa González. A unique cluster of roo insertions in the promoter region of a stress response gene in *Drosophila melanogaster*. *Mobile DNA*, 10(10), 03 2019.

[5] Dmitri A. Petrov, Anna-Sophie Fiston-Lavier, Mikhail Lipatov, Kapa Lenkov, and Josefa González. Population Genomics of Transposable Elements in *Drosophila melanogaster*. *Molecular Biology and Evolution*, 28(5):1633–1644, 12 2010.

Implementing a Text Mining Service Offer on the Migale Bioinformatics Platform

Mouhamadou Ba[1,2], Véronique Martin[1,2], Olivier Rué[1,2], Sophie Schbath[1,2], Valérie Vidal[1,2], Valentin Loux[1,2]

[1] Université Paris-Saclay, INRAE, MaIAGE, Jouy-en-Josas, France
[2] Université Paris-Saclay, INRAE, BioinfOmics, MIGALE Bioinformatics Facility, Jouy-en-Josas 78350, France

Corresponding Author: mouhamadou.ba@inrae.fr

The large and growing amount of textual data (scientific articles, reports, database fields, etc.) in the scientific domains, particularly in Life Sciences, are a valuable source of information and knowledge. The data are however in practice inaccessible with traditional review techniques and remain largely under-utilized [1]. Recent advances in the text and data mining (TDM) technologies, the change in the legal and regulatory aspects, and the development of the scientific service infrastructures are paving the way for new solutions to discover, harness, integrate and learn from the textual data.

The Migale Bioinformatics facility (https://migale.inrae.fr), taking advantage of that context, is working to complete its service offer [2] with text mining services to make access to bioinformatics research communities text mining solutions for analyzing and extracting value from textual data. The development of the offer will be based on the existing Migale infrastructure and on technologies developed by the TDM communities [3,4]. Migale has already added text mining to its existing Data Analysis Service (https://migale.inrae.fr/ask-data-analysis) in order to propose collaboration to users who want to address information extraction from texts in the microbiology fields (creation of thematic corpora from sources like PubMed, extraction of named entities such as genes, phenotypes, metabolites; and classification of parts of texts).

Migale has added a training module entitled "Introduction to Text Mining with AlvisNLP" to its "Bioinformatics by practicing" cycle (https://migale.inrae.fr/trainings). The module is open to (bio)informaticians who want to benefit from theoretical and practical skills in the analysis of textual data. It addresses Named Entity Recognition (REN) methods through use cases in biology (recognition of genes, proteins or habitats of bacteria, etc.).

Migale also ensures the deployment and management of several web applications (Semantic Search Engines, Annotation Editors, Terminology Editor) for partners. We develop on-demand APIs that encapsulate specialized text mining process focusing on well-defined use cases within projects (e.g., we propose in project TIERS-ESV an API for the automatic extraction of information about pest organisms to facilitate animal health monitoring).

Acknowledgements

The work is done in close collaboration with the INRAE Bibliome Research Team (Bibliome, INRAE, 2022. https://maiage.inrae.fr/fr/bibliome). The service offer relies today mainly on text mining resources they have developed. Special thanks to R. Bossy (https://orcid.org/0000-0001-6652-9319) who takes part to the development and training activities; and maintains some of the precious text mining tools we use.

References

1. Chaix E., Deléger L., Bossy R., Nédellec C. *Text Mining Tools for Extracting Information about Microbial Biodiversity*, In Food Microbiology, 2018.
2. Loux Valentin et al. *The Migale Bioinformatique facility*. In Proceedings of JOBIM, 2021
3. Bossy R., Deléger L., Chaix E., Ba M., Nédellec C. *Bacteria Biotope at BioNLP Open Shared Tasks 2019,* In Proceedings of the 5th workshop on BioNLP open shared tasks, 2019
4. Ba M., Bossy R. *Interoperability of corpus processing workflow engines: the case of AlvisNLP/ML in OpenMinTed*, In Proceedings of the Workshop on Cross-Platform Text Mining and Natural Language Processing Interoperability at LREC, 2016

Generation of a genome-wide 3D RNA profile of the catshark habenulae

Hélène MAYEUR[1], Léo MICHEL[1], Sébastien DEJEAN[2], Patrick BLADER[3] Ronan LAGADEC[1], Sylvie MAZAN[1]

[1] CNRS, Sorbonne Université, UMR7232-Biologie Intégrative des Organismes Marins (BIOM), Observatoire Océanologique, Banyuls sur Mer, France
[2] Institut de Mathématiques de Toulouse, Université de Toulouse, CNRS, UPS, UMR 5219, Toulouse, France
[3] Centre de Biologie Intégrative (CBI, FR 3743), Université de Toulouse, CNRS, UPS, France

Corresponding author: helene.mayeur@obs-banyuls.fr

Habenulae are bilateral epithalamic structures found in all vertebrates and involved in the integration of sensory perceptions and the regulation of adaptive behaviors. A remarkable feature of these structures is that they display asymmetries between the left and the right sides in many vertebrate species. Their subdomain organization substantially differs between the mouse and zebrafish, and how it evolved across vertebrates remains poorly known. In order to gain insight into the mode of evolution of habenulae across jawed vertebrates, we focused on a chondrichthyan species, the catshark *Scyliorhinus canicula*. This species harbors a key phylogenetic position as a member of the sister group of osteichthyans and has retained putative ancestral characteristics of habenular subdomain organization, including asymmetries lost in the mouse and the zebrafish (Lanoizelet, Michel, Lagadec, Mayeur *et al.* unpublished results).

Our experimental strategy consisted in generating a genome-wide 3D RNA profile of the habenulae, using RNA tomography, a section-based technique that yields spatial resolution to RNA-seq data [1, 2]. To do so, we constructed and sequenced barcoded cDNA Illumina libraries starting from RNA extracted from serial sections along transverse, horizontal and sagittal planes using the Cel-Seq2 single-cell RNA-seq protocol, mapped the reads on a gene model reference and projected 1D read counts onto a 3D digital model of the habenulae by iterative proportional fitting. We thus generated a genome-wide 3D RNA profile of the catshark habenulae, containing quantified expression data for about 20000 protein-coding genes in 95000 voxels.

The digital signals obtained for known habenula subdomain markers reproduce the broad characteristics of their in situ hybridization profiles, which validates this model. We used correlation and spatial autocorrelation analyses to identify (1) novel candidate markers for the habenula subdomains previously characterized and (2) novel organ subterritories. In situ hybridizations are in progress to validate these in silico data but preliminary results already confirm the potential of this approach to refine our understanding of habenular organisation and to expand the repertoire of signature markers of habenular subdomains. The resulting characterization of catshark habenulae should provide a reference allowing exhaustive comparisons across vertebrates and the generation of a comprehensive scenario for habenular evolution in the taxon.

References

1. F. Kruse, J. P. Junker, A. van Oudenaarden and J. Bakkers. Tomo-seq: a method to obtain genome-wide expression data with spatial resolution. Methods in Cell Biology. 135:299-307, 2016.
2. Hélène Mayeur, Maxence Lanoizelet, Aurélie Quillien, Arnaud Menuet, Léo Michel, Kyle John Martin, Sébastien Dejean, Patrick Blader, Sylvie Mazan and Ronan Lagadec. When Bigger Is Better: 3D RNA Profiling of the Developing Head in the Catshark *Scyliorhinus canicula*. *Front Cell Dev Biol.* 9:744982, 2021.

Novel adversarial autoencoders to simulate human genomic data for clinical research

Callum Burnard[1], William Ritchie[1] and Alban Mancheron[2]

[1] IGH, 141 rue de la Cardonille, 34094, Montpellier, France

[2] LIRMM, 161 rue Ada, 34095, Montpellier, France

Corresponding Author: callum.burnard@igh.cnrs.fr

1. Context

Complete human genomes available in full online are quite rare, with the current largest repository of publicly available human genomes being that of the 1000 Genomes Project [1], counting data for just over 2500 people in its final phase. This is understandable: allowing the world to download your genome represents a major privacy risk. However, they represent a very useful resource for researchers, as all studies need large sample sizes for statistical relevance. Applying machine learning methods to these datasets would allow users to produce synthetic genomes that still follow the implicit rules of the genomes it has learned from, but this poses a challenge due to the scale of data to analyze. Previous work has applied generative neural networks (GNNs) to subsections of the human genome [2], demonstrating that the principle is sound, but complex to apply at the necessary scale. Their findings indicate that a GNN is able to generate new sample consisting of up to 10,000 SNPs with a diversity similar to that of the input dataset.

2. Methods

In this project, we separate publicly available human genomes into subsections separated by recombination hotspots, based off of a map of recombination likelihood [3], then apply dimensionality reduction methods to each section. We use deep autoencoders, which act as an extension of our GNN, but are able to be trained individually to save on computational resources required. This allows a GNN to simulate genomic data on a much larger scale in two steps: first by producing compressed data, then by feeding that data into the corresponding autoencoder for expansion.

3. Results

Preliminary results show that autoencoders are well suited to compress genomic data by at least an order of magnitude while still maintaining an accuracy of 95% or better. Our first Generative Adversarial Network [4] models are able to learn patterns within this latent space and reproduce realistic samples based on this input. Applying them to the entirety of chromosome 1 causes issues in the leaning process, specifically mode collapse, even when using Wasserstein loss to reduce the likelihood of this occurring [5]. Currently, we are aiming to test the limits of this method to see just how many mutations we are able to simulate, and to what degree of diversity and realism.

Acknowledgements

The authors would like to thank la Ligue Contre le Cancer for providing funding for C. Burnard's PhD project. The authors would also like to thank the Genotoul platform for their high-performance computing services and data storage.

References

1. Ernesto Lowy-Gallego, et al. *Variant calling on the GRCh38 assembly with the data from phase three of the 1000 Genomes Project*. Wellcome Open Research, 2019.
2. Burak Yelmen, et al. *Creating artificial human genomes using generative neural networks*. PLOS Genetics, 2021.
3. Bjarni V Halldorsson, et al. *Characterizing mutagenic effects of recombination through a sequence-level genetic map*. Science, 2019.
4. Ian J Goodfellow, et al. *Generative Adversarial Nets*. NEURIPS Proceedings, 2014.
5. Martin Arjovsky, Soumith Chintala, Léon Bottou. *Wasserstein GAN*. arXiv, 2017.

Développement d'une méthode quantitative d'analyse de données RNA-Seq pour l'étude des variations *in vivo* des états de phosphorylation en 5' des ARN et application chez *Staphylococcus aureus*

Tomas CAETANO[1], Yves QUENTIN[1] Peter REDDER[2], Gwennaele FICHANT[1], Roland BARRIOT[1]

[1]Génomique des systèmes intégrés [2]Flux des ARN et régulation génique chez *Staphylococcus aureus* - Laboratoire de Microbiologie et de Génétique Moléculaires, Centre de Biologie Intégrative, Université de Toulouse, France

Auteur correspondant : tomas.caetano@univ-tlse3.fr

La dégradation de l'ARN joue un rôle essentiel dans la régulation des gènes bactériens. Chez *Staphylococcus aureus*, cette dégradation est principalement effectuée par les ribonucléases RNase J. Ces enzymes possèdent à la fois une activité endo – et 5'-exoribonucléasique. Cette dernière semble être contrôlée par plusieurs ARN pyrophosphohydrolases (RPP) différentes, qui modifient de manière post-transcriptionelle l'état de phosphorylation des extrémités 5' des ARN, de la forme 5'PPP à la forme 5'P, et permettent leur dégradation par la RNase J[1-3]. Nous avons développé une nouvelle méthode de transcriptomique permettant de distinguer l'extrémité 5'PPP (non modifiée après transcription) des extrémités 5'P et 5'PP des ARN et de quantifier leurs proportions pour chaque espèce d'ARN : PQ-EMOTE pour *Phosphorylation Quantification Exact Mapping Of Transcriptome Ends*.
Nous présentons ici, une méthode d'analyse unique et innovante permettant de traiter les données générées par PQ-EMOTE. Par rapport à l'analyse différentielle de gènes par RNA-Seq qui regroupe généralement le comptage des lectures par régions génomiques, notre méthode permet d'obtenir uniquement des lectures dont l'extrémité 5' correspond à celles observées *in vivo*, et ainsi, seule la 1re position génomique du transcrit aligné sur le génome de référence est prise en compte (position qui correspond au +1 de transcription). Un traitement spécifique des données a été élaboré pour la quantification des lectures à la position près et leur normalisation. Une autre particularité dans l'analyse provient du fait que nous cherchons à détecter et identifier des variations de proportions (d'états de phosphorylation de l'extrémité 5' pour un même +1 de transcription) et non une variation d'abondance de transcrits pour une région génomique.

Nous présentons des résultats préliminaires de l'application de cette méthode pour identifier, chez *S. aureus,* les protéines ayant une activité RPP. Bien que centrale dans le métabolisme de l'ARN, aucune protéine réalisant cette fonction n'est connue dans cette espèce. Jusqu'à maintenant, les RPP identifiées chez d'autres bactéries appartiennent à la superfamille Nudix dont les membres présentent peu de similarité de séquence, même ceux ayant les mêmes fonctions[4], rendant leur identification par correspondance entre espèces peu fiable. De plus, ces Nudix sont souvent décrites comme possédant plusieurs fonctions et il est donc possible que plusieurs RPP soient présentes chez *S. aureus* et les effets de cross-talk doivent donc être pris en compte dans l'analyse. Les résultats présentés concernent plusieurs jeux de données correspondants à des comparaisons entre la souche sauvage et des mutants de délétion simple, double ou triple de quatre des cinq gènes codant pour les protéines RPP candidates.

[1] N. Mathy, L. Bénard, O. Pellegrini, R. Daou, T. Wen, et C. Condon, « 5'-to-3' exoribonuclease activity in bacteria: role of RNase J1 in rRNA maturation and 5' stability of mRNA », *Cell*, vol. 129, nº 4, p. 681-692, mai 2007, doi: 10.1016/j.cell.2007.02.051.

[2] S. Hausmann, V. A. Guimarães, D. Garcin, N. Baumann, P. Linder, et P. Redder, « Both exo- and endo-nucleolytic activities of RNase J1 from Staphylococcus aureus are manganese dependent and active on triphosphorylated 5′-ends », *RNA Biol.*, vol. 14, nº 10, p. 1431-1443, oct. 2017, doi: 10.1080/15476286.2017.1300223.

[3] S. Even, « Ribonucleases J1 and J2: two novel endoribonucleases in B.subtilis with functional homology to E.coli RNase E », *Nucleic Acids Res.*, vol. 33, nº 7, p. 2141-2152, avril 2005, doi: 10.1093/nar/gki505.

[4] A. G. McLennan, « Substrate ambiguity among the nudix hydrolases: biologically significant, evolutionary remnant, or both? », *Cell. Mol. Life Sci. CMLS*, vol. 70, nº 3, p. 373-385, févr. 2013, doi: 10.1007/s00018-012-1210-3.

Predicting gene regulation through co-occurrence and evolutionary conservation of transcription factor binding sites

Laura TURCHI[1,2], Antoine FRENOY[2], Nicolas THIERRY-MIEG[2], Romain BLANC-MATHIEU[1] and François PARCY[1]

[1] Laboratoire Physiologie Cellulaire et Végétale, Univ. Grenoble Alpes, CNRS, CEA, INRAE, IRIG-DBSCI-LPCV, 17 avenue des martyrs, F-38054, Grenoble, France

[2] TIMC, Univ. Grenoble Alpes, CNRS, UMR5525, Rond-Point de la Croix de Vie, 38706, La Tronche, France

Corresponding Author: francois.parcy@cea.fr

Transcription factors (TFs) are DNA-binding proteins that play a crucial role in gene regulation. To control gene expression TFs bind specific short stretches of DNA called transcription factor binding sites (TFBS). However, the intrinsic and predictable binding affinity of a TF for its TFBS does not always translate into *in vivo* binding and to a transcriptional effect. Thus, models based on motif recognition can be poor predictors of 'functional TFBS' (i.e. TFBS that are bound and regulatory *in vivo*).

Functionally important TFBS tend to be evolutionarily conserved [1], and multiple TFs can bind to the same regulatory region to fine-tune gene expression [2]. We aimed to characterize and build a machine learning model to predict TF-dependent gene regulation, leveraging information about (i) co-occurrence of multiple TFs close to a TF of reference, and (ii) evolutionary conservation of the putative TFBS. We applied this approach on LEAFY (LFY), a plant-specific TF playing a crucial role in floral development. LFY is highly conserved in sequence and binding specificity [3], and the availability of binding and expression data makes it possible to train a machine learning model and validate predictions. I will show preliminary results obtained with this approach to predict gene regulation by LFY.

Acknowledgements

This project was funded by CNRS 80|Prime.

References

1. K. R. Nitta *et al.*, « Conservation of transcription factor binding specificities across 600 million years of bilateria evolution », *eLife*, vol. 4, p. e04837, mars 2015, doi: 10.7554/eLife.04837.
2. F. Reiter, S. Wienerroither, et A. Stark, « Combinatorial function of transcription factors and cofactors », *Curr. Opin. Genet. Dev.*, vol. 43, p. 73-81, avr. 2017, doi: 10.1016/j.gde.2016.12.007.
3. C. Sayou *et al.*, « A Promiscuous Intermediate Underlies the Evolution of LEAFY DNA Binding Specificity », *Science*, vol. 343, n° 6171, p. 645-648, févr. 2014, doi: 10.1126/science.1248229.

Identification of epimutations in rare diseases from a single patient perspective

Robin Grolaux[1], Alexis Hardy[1] and Matthieu Defrance[1]
IB square, Université Libre de Bruxelles, Belgium

Corresponding author: robin.grolaux@ulb.be

1 Introduction

DNA methylation (DNAm) plays an important role in cell biology, most notably for tissue specific regulation of gene expression, other roles include X-chromosome inactivation, regulation of splice-junctions and genomic imprinting [1,2]. Changes in DNAm (i.e. epimutations) arise either from stochastic errors in the establishment or maintenance of a methylation state by the DNMTs protein family (i.e. primary epimutations), or following a change in the DNA sequence (i.e. secondary epimutations) [3]. Both primary and secondary epimutations are found in patients suffering from rare diseases, a worldwide public health issue estimated to affect between 260 and 445 million people [4]. Identification of primary epimutations can lead to a direct diagnosis, such is the case in imprinting disorders [5], rare cases of cancer [6,7] and some neurodevelopmental diseases [8]. Secondary epimutations in rare diseases have gained popularity as a surrogate to the detection of sequence variants in the diagnosis process, as they are easier to detect. Therefore, identification of variations in DNAm plays an important role in the understanding of the aetiology of those diseases as well as in the clinical diagnosis process. Canonical pipelines for the detection of epivariants based on methylation-array technologies rely on case-control group comparisons. However, in the context of rare diseases and multi-locus imprinting disorders, small cohorts and inter-patients' heterogeneity prevent the use of those tools. Therefore, there is a need to provide a comprehensible and statistically robust pipeline for clinicians to perform analyses at the single patient level as well as characterise how different parameters may influence epivariants detection. This poster describes a statistical method to detect differentially methylated regions in correlated datasets based on the z-score and the empirical Brown aggregation method from a single patient perspective. It further provides a characterisation of how the chosen parameters may influence epivariants detection. We generated semi-simulated data based on a public control population of 521 samples. This enabled us to evaluate how control population, effect and region size affect the performance of epivariants detection, in order to define the optimal parameters of the method. Finally, we validated the detection of pathological methylation events in patients suffering from rare multi-locus imprinting disturbances and showed how this method is complementary to the validation of clinical diagnosis.

References

[1] Maxim V. C. Greenberg and Deborah Bourc'his. The diverse roles of dna methylation in mammalian development and disease. *Nature Reviews Molecular Cell Biology*, 20(10):590–607, 2019.

[2] Galit Lev Maor, Ahuvi Yearim, and Gil Ast. The alternative role of dna methylation in splicing regulation. *Trends in Genetics*, 31(5):274–280, 2015.

[3] B. Horsthemke. *Epimutations in Human Disease*, pages 45–59. Springer Berlin Heidelberg, Berlin, Heidelberg, 2006.

[4] S. Nguengang Wakap, D. M. Lambert, A. Olry, C. Rodwell, C. Gueydan, V. Lanneau, D. Murphy, Y. Le Cam, and A. Rath. Estimating cumulative point prevalence of rare diseases: analysis of the Orphanet database. *Eur J Hum Genet*, 28(2):165–173, 02 2020.

[5] Thomas Eggermann, Guiomar Perez de Nanclares, Eamonn R. Maher, I. Karen Temple, Zeynep Tümer, David Monk, Deborah J. G. Mackay, Karen Grønskov, Andrea Riccio, Agnès Linglart, and Irène Netchine. Imprinting disorders: a group of congenital disorders with overlapping patterns of molecular changes affecting imprinted loci. *Clinical Epigenetics*, 7(1):123, 2015.

[6] M P Hitchins and R L Ward. Constitutional (germline) mlh1 epimutation as an aetiological mechanism for hereditary non-polyposis colorectal cancer. *Journal of Medical Genetics*, 46(12):793–802, 2009.

[7] Estela Dámaso, Adela Castillejo, María del Mar Arias, Julia Canet-Hermida, Matilde Navarro, Jesús del Valle, Olga Campos, Anna Fernández, Fátima Marín, Daniela Turchetti, Juan de Dios García-Díaz, Conxi

Lázaro, Maurizio Genuardi, Daniel Rueda, Ángel Alonso, Jose Luis Soto, Megan Hitchins, Marta Pineda, and Gabriel Capellá. Primary constitutional mlh1 epimutations: a focal epigeneticevent. *British Journal of Cancer*, 119(8):978–987, 2018.

[8] Mafalda Barbosa, Ricky S. Joshi, Paras Garg, Alejandro Martin-Trujillo, Nihir Patel, Bharati Jadhav, Corey T. Watson, William Gibson, Kelsey Chetnik, Chloe Tessereau, Hui Mei, Silvia De Rubeis, Jennifer Reichert, Fatima Lopes, Lisenka E. L. M. Vissers, Tjitske Kleefstra, Dorothy E. Grice, Lisa Edelmann, Gabriela Soares, Patricia Maciel, Han G. Brunner, Joseph D. Buxbaum, Bruce D. Gelb, and Andrew J. Sharp. Identification of rare de novo epigenetic variations in congenital disorders. *Nature Communications*, 9(1):2064, 2018.

Dynamic genes network inference and very short time series. How repeated acoustic stimuli affect plant immunity?

Khaoula Hadj-Amor[1], Adelin Barbacci[2] and Frédérick Garcia[1]

[1] unité de Mathématiques et Informatique Appliquées de Toulouse (MIAT) , Institut National de Recherche pour l'Agriculture, l'alimentation et l'Environnement (INRAe), 24 Chemin de Borde-Rouge, 31326, Castanet-Tolosan, FRANCE

[2] Laboratoire des Interactions Plantes Microorganismes et Environnement (LIPME) , Institut National de Recherche pour l'Agriculture, l'alimentation et l'Environnement (INRAe) - Centre National de la Recherche Scientifique (CNRS), 24 Chemin de Borde Rouge, 31326, Castanet-Tolosan, FRANCE

Corresponding author: khaoula.hadj-amor@inrae.fr

1 Introduction

RNA-sequencing methods are central to biology and the basis of many breakthroughs in understanding the transcriptome. Current sequencing techniques offer the opportunity to understand life in its spatial and temporal aspects. However, the high dimensionality of RNA-seq data limits the understanding of complex dynamic relations between genes and mathematical methods must be implemented to ease the comprehension of the underpinning biology. Dynamic network inference methods are methods of choice to integrate the complexity of time-dependant transcriptomic data [1]. These methods produce networks in which nodes are genes and an edge between two genes represents an action (up or down-regulation) of a gene on another over time. Although, most network inference methods are limited by the large dimension of RNA-seq data, in which the number of observations is always much lower than the number of measured gene expressions. Hence, the data dimension is frequently reduced by performing the inference on a small number of genes, picked *a priori*. In this work, we propose another strategy, which consists of grouping genes with similar expression dynamics to form a clustered gene network. More precisely, our clustered gene network inference strategy is a two-step method (clustering then network inference). The clustering step is based on the analysis of sign variations in order to make it suitable for biology-associated very-short time series. Our clustered gene network inference strategy is illustrated on original transcriptomic data associated with the priming of plant immune resistance by repeated acoustic stimuli (RAS) over time.

2 Results

Arabidopsis thaliana plants were exposed to acoustic stimuli (1KHz, 100dB) for 3 hours per day for 0 to 8 days prior to the infection by the necrotrophic fungus *Sclerotinia sclerotiorum*. After 3 RAS, plants exhibited a 12% significant gain of resistance to the disease. To better understand the transcriptomic reprogramming underpinning this huge gain of resistance, RNA sequencing was performed on healthy plants prior to the infection after 0, 1, 3, 8 RAS. The differential analysis led to identifying 9554 RAS modulated genes. We restricted our analysis to these genes. We benchmarked our sign variations clustering methods with standard methods (k-means for functional data, hierarchical for time series data, k-means longitudinal data, k-means and hierarchical clustering). The sign variations methods consist of 2 steps. Genes with similar temporal variations were first grouped together. Each group was next split by the k-means clustering method into subgroups whose number is determined by the silhouette coefficient. The clustering provided by the sign variations method exhibited the lowest distance within clusters and the lowest Davies-Bouldin index with a reduced computational time. We also found that the clustering method impacted the mean squared prediction error of VAR(1) inferred network [2]. The sign variations clustering method led to a reduced prediction error compared to other clustering methods. The sign variations clustering appears suitable for network inference based on very-short time series of gene expression.

References

[1] Christopher A Penfold and David L Wild. How to infer gene networks from expression profiles, revisited. *Interface focus*, 1(6):857–870, 2011.

[2] George Michailidis and Florence d'Alché Buc. Autoregressive models for gene regulatory network inference: Sparsity, stability and causality issues. *Mathematical biosciences*, 246(2):326–334, 2013.

A general framework for classifying genomic sequences with Transformers: Application to gene annotation.

Matthias LORTHIOIS[1,2], Édouard CADIEU[1], Aurore BESSON[1], Catherine ANDRÉ[1], Benoit HÉDAN[1], Christophe HITTE[1], Thomas DERRIEN[1]

[1]Univ Rennes, CNRS, IGDR – UMR6290, F-35000 Rennes, France.
[2]Univ Rouen-Normandy, Master of Bioinformatics, Modelling and Statistics, F-76000 Rouen, France.

Corresponding Authors: matthias.lorthiois@univ-rennes1.fr, tderrien@univ-rennes1.fr

The development of transcriptome sequencing (RNASeq) has allowed to rapidly identify and quantify all RNA molecules in a cell type or a tissue. This shed light on the pervasive transcription of the genomes where functionally important protein-coding (mRNAs) and long non-coding RNAs (lncRNAs) need to be distinguished from transcriptional noise [1]. Deep learning (DL) strategies using either convolutional neural networks, recurrent neural networks or more recently Encoders with self-attention mechanism (*e.g.* Transformers), have been successfully applied to various genomic classification tasks such as chromatin accessibility or regulatory sequence classification. Here, we develop a general and easy-to-use framework, called `TransforKmers`, which facilitates the building, training and utilization of Transformers models, as provided by HuggingFace [2], for genomic sequence classification. More specifically, all essential steps are included in `TransforKmers`: the creation of a pretraining dataset from any given genomic sequences, the configuration of customizable tokenizers, the pretraining of the model and its finetuning with labeled sequences, the test and evaluation of the model and finally, the inference on real sequences. To illustrate the usability of our framework, we applied it to classify novel transcription start sites (TSSs) reconstructed from 6 long-read RNASeq (LR-RNAseq) targeting canine transcriptomes. Hence, we pretrained a Transformer architecture inspired by BERT [3] with 2 million DNA genomic sequences of length 512 nt. We then fine-tuned the model with a positive set of 94,100 TSSs extracted from the canFam4 NCBI annotation and a corresponding negative set composed of both random intergenic and real sequences randomly shuffled. Our model achieved high performance with F1-score of 0.912 and accuracy of 0.914 on the test set. Interestingly, we highlighted that TSSs classified as true positives are enriched for CAGE (Cap Analysis of Gene Expression) peaks (χ^2, p-value $=2.2.10^{-16}$) strengthening the accuracy of the 5'-end predictions. We then applied our method on a real dataset of novel canine genes (both mRNAs and lncRNAs) that were annotated by our previously published pipeline `ANNEXA` [4]. We showed that 43,9% (2274/5183) of the novel TSSs (*i.e.* not annotated in the reference transcriptome) were classified as positives. Given that LR-RNASeq protocols start from the 3'end of the transcripts (polyA tail), these transcripts can be considered full-length, thus paving the way for further functional validations. Altogether, we developed a novel DL-based framework which facilitates the processing and classification of biological sequences based on Transformer architectures. This tool has been integrated as a new module in our ANNEXA pipeline and is also freely available at GitHub: https://github.com/mlorthiois/TransforKmers.

Acknowledgements

Authors would like to warmly thank the Bioinformatics Genouest platform (https://www.genouest.org) for providing the required infrastructure for this work and especially the (geno)gpus.

References

[1] M. Pertea *et al.*, "CHESS: a new human gene catalog curated from thousands of large-scale RNA sequencing experiments reveals extensive transcriptional noise," *Genome Biol*, Dec. 2018, doi: 10.1186/s13059-018-1590-2.

[2] T. Wolf *et al.*, "HuggingFace's Transformers: State-of-the-art Natural Language Processing," arXiv, arXiv:1910.03771, Jul. 2020. doi: 10.48550/arXiv.1910.03771.

[3] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding." arXiv, May 24, 2019. Available: http://arxiv.org/abs/1810.04805

[4] Lorthiois M. *et al.*, "ANNEXA: Analysis of Nanopore transcriptomic data with Nextflow for Extended Annotation" JOBIM-Paris, 2021.

Using contrast to study RNA transcripts co-maturations

Benjamin VACUS[1,2], Arnaud LIEHRMANN[1,2,3], Guillem RIGAILL[1,2,3], Benoit CASTANDET[1,2], and Etienne DELANNOY[1,2]

[1] Institute of Plant Sciences Paris-Saclay (IPS2), Université Paris-Saclay, CNRS, INRAE, Université Evry, 91405, Orsay, France

[2] Institute of Plant Sciences Paris-Saclay (IPS2), Université Paris Cité, CNRS, INRAE, 91405, Orsay, France

[3] Laboratoire de Mathématiques et de Modélisation d'Evry (LaMME), Université d'Evry-Val-d'Essonne, UMR CNRS 8071, ENSIIE, USC INRAE, 91000, Evry, France

Corresponding Author: benjamin.vacus@universite-paris-saclay.fr

Background. Plant cells harbor three different genomes: in addition to the nucleus, genes are found in other organelles, such as mitochondria or chloroplasts. Inside chloroplasts, RNA transcripts undergo a complex set of maturation events including *splicing*, *editing* – the transformation of one base into another – and *processing of their extremities* [1]. We recently took advantage of the Nanopore sequencing technology to study the dependencies between two different maturation events on the same transcript in *Arabidopsis thaliana* [2]. This long read technology allowed us to jointly monitor events that are sometimes separated by several thousand bases. Dependencies were measured using a Fisher test, thus ignoring several important features of the sequencing data as count dispersion and replicates intrinsic variability.

R pipeline. Contingency tables containing the number of reads in each maturation state for every pair of maturation events were used as input to a conventional gene differential expression analysis pipeline. R DESeq2 package [3] was used for computing contrasts and applying a Wald test. In the simple case of studying the dependence of pairwise maturation events in one biological condition, testing the contrast is reduced to testing the 2nd order interaction term of the generalized linear model. Adjusted p-values led to the estimation of dependency between events in each pair.

Empirical Results. We showed that the contrast method was able to recover most of the dependencies previously found with the Fisher test, with a similar ranking between all pairs of events. Differences between the two methods mainly lie in the most ambiguous cases, i.e., when the p-value approaches 5%.

Discussion. In the two-events case, the contrast method proved effective. Dependencies obtained using the Fisher test and rejected by the contrast method might be considered false positives since this new method estimates the counts' dispersion more accurately using the binomial-negative hypotheses and models the specificity of each replicate.

Further Work. We plan to adapt the method to the study of dependencies between more than two events, and to integrate a second biological condition or the maturation event processing of the ends in the pipeline.

References

1. Stern, David B., Michel Goldschmidt-Clermont, and Maureen R. Hanson. Chloroplast RNA metabolism. *Annual review of plant biology 61*: 125-155, 2010.
2. Guilcher, M., Liehrmann, A., Seyman, C., Blein, T., Rigaill, G., Castandet, B., & Delannoy, E.. Full length transcriptome highlights the coordination of plastid transcript processing. International journal of molecular sciences, 22(20), 11297, 2021.
3. Love, Michael I., Wolfgang Huber, and Simon Anders. "Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2." Genome biology 15.12: 1-21, 2014.

Deep learning approaches as scoring methods for protein-protein rigid body docking.

Helene BRET[1], Jessica ANDREANI[1] and Raphael GUEROIS[1]

[1] Université Paris-Saclay, CEA, CNRS, Institute for Integrative Biology of the Cell (I2BC), 91198, Gif-sur-Yvette, France

Corresponding Author: helene.bret@cea.fr

Proteins perform most of their functions by interacting with other proteins. Studying protein complexes is crucial to better understand the interaction mechanisms at the molecular level. Determining experimental structures is a time consuming and costly process and is not always possible. Computational methods for structure prediction have therefore been developed to identify in silico the most likely conformations of the bound partners of a complex. Rigid body docking is a two-step approach: the sampling step and the scoring step. During sampling, a large set of possible conformations is generated. Conformations are then ranked to identify the most likely poses (scoring step).

Several scoring methods have been proposed based on physics, statistical and evolutionary information as in the consensus score implemented in the server InterEvDock3 [1] developed by our team. In parallel, deep learning approaches have proven to be extremely powerful to study the structure of biological objects, by extracting a signal from a covariation map as in ComplexContact [2] or TrRosetta [3], or more recently by analyzing protein sequences with the Transformers technology in the successful AlphaFold2 method [4].

In this work we present a deep learning approach analyzing proteins at the residue level. Inputs of our model include geometric and sequence information for one residue and its environment. The model evaluates the adequacy between the residue and its context. A more global representation of the whole complex structures is being developed, using graph representation and graph convolutions.

References

1. Quignot, C., Postic, G., Bret, H., Rey, J., Granger, P., Murail, S., ... & Guerois, R. (2021). InterEvDock3: a combined template-based and free docking server with increased performance through explicit modeling of complex homologs and integration of covariation-based contact maps. *Nucleic Acids Research*, 49(W1), W277-W284.
2. Zeng, H., Wang, S., Zhou, T., Zhao, F., Li, X., Wu, Q., & Xu, J. (2018). ComplexContact: a web server for inter-protein contact prediction using deep learning. *Nucleic acids research*, 46(W1), W432-W437.
3. Yang, J., Anishchenko, I., Park, H., Peng, Z., Ovchinnikov, S., & Baker, D. (2020). Improved protein structure prediction using predicted interresidue orientations. *Proceedings of the National Academy of Sciences*, 117(3), 1496-1503.
4. Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., ... & Hassabis, D. (2021). Highly accurate protein structure prediction with AlphaFold. *Nature*, 596(7873), 583-589.

Robust deconvolution of transcriptomic samples using the gene covariance structure

Bastien CHASSAGNOL[1,2,3], Pierre-Henri WUILLEMIN[1], Gregory NUEL[2] and Etienne BECHT[3]
[1] LIP6 (Laboratoire d'Informatique Paris 6), 4 Place Jussieu, 75005, Paris, FRANCE
[2] LPSM (Laboratoire de Probabilités, Statistiques et Modélisation), 4 Place Jussieu, 75005, Paris, FRANCE
[3] Les Laboratoires Servier, 50 Rue Carnot, 92150, Suresnes, FRANCE

Corresponding author: bastien.chassagnol@upmc.fr

Transcriptomic analyses have increasingly contributed to our understanding of the intricate biological processes involved in the emergence of auto-immune diseases or tumour-promoting environments. However, classical bulk analyses ignore the intrinsic complexity of biological samples, by averaging measurements over multiple distinct cell populations. It is therefore unclear whether a change in the gene expression between samples results from a variation of the cell type proportions, from an environmental signal or a mutation [1].

To remove this ambiguity, deconvolution algorithms can estimate the proportions of cell populations from a bulk transcriptome using the reference transcriptome of purified cell populations. Traditionally, most approaches, including the gold standard CIBERSORT algorithm [2], retrieve the cell proportions of a mixture assuming the linear assumption that each gene expression is the sum of each cell population's contribution weighted by their corresponding relative frequency in the sample.

However, none of these methods account for the transcriptomic covariance structure and address the crucial problem of co-expression between distinct genes. The first goal of our project aims at studying the impact of correlation structures in the quality of the estimation performed by canonical deconvolution algorithms that assume *iid* distributions between the genes and use a fixed averaged expression profile for each cell type. The transcriptomic pathways were learnt from publicly purified cell data only, hypothesising that the network structure was sparse. Direct connections between the genes are represented for each population by non-zeros entries, learnt by plugging in the *MLE* covariance estimate, with zeros inputs shrunk by the gLasso algorithm [3,4].

Then, we develop a new deconvolution method that model each purified cellular expression profile as a multivariate Gaussian distribution [5], whose covariance parameter is the *plugged-in* estimate learnt beforehand to reconstitute the bulk profile. Next, we will optimise the estimation of the cellular expression profiles, by determining the MLE optimising the associated convolution of density functions of purified multivariate Gaussian transcriptomic profiles. Finally, we will compare our method to standard deconvolution algorithms, showing its interest to supply estimates more faithful to the biological reality.

References

[1] Shai S. Shen-Orr and Renaud Gaujoux. Computational deconvolution: Extracting cell type-specific information from heterogeneous samples. *Current Opinion in Immunology*, 5, 2013.

[2] Aaron Newman, Chih Liu, Michael Green, Andrew Gentles, Weiguo Feng, Yue Xu, Chuong Hoang, Maximilian Diehn, and Ash Alizadeh. Robust enumeration of cell subsets from tissue expression profiles. *Nature methods*, 12, 2015.

[3] Jerome Friedman, Trevor Hastie, and Robert Tibshirani. Sparse inverse covariance estimation with the graphical lasso. *Biostatistics*, 3:432–441, 2008.

[4] Joachim Dahl, Vwani Roychowdhury, and Lieven Vandenberghe. Maximum likelihood estimation of gaussian graphical models: Numerical implementation and topology selection. *Journal of Multivariate Analysis*, 2005.

[5] Daphne Koller and Nir Friedman. *Probabilistic Graphical Models: Principles and Techniques*. MIT Press, 2009.

AOP-helpFinder : a tool for exploration of the literature to support adverse outcome pathways development

Thomas Jaylet[1], Florence Jornod[1] and Karine Audouze[1]

[1] Université Paris Cité, T3S - Inserm UMRS 1124, 75006, Paris, France

Corresponding Author: thomas.jaylet@parisdescartes.fr

Abstract

The Adverse Outcome Pathway (AOP) is a conceptual framework proposed by Ankley et al. (2010) that was developed to address some of the toxicological and ecotoxicological challenges of the 21st century [1]. AOPs allow to organize toxicological and ecotoxicological data as a linear combination of biological events initiated by exposure to a stressor (e.g., pollutants, chemicals, stresses). An AOP always starts with a molecular initiating event (MIE), progresses via a cascade of key events (KEs) through different levels of biological organization (cells, tissues, organs) to the appearance of an Adverse Outcome (AO). Biological events (MIE, KE, AO) are not exclusive to one AOP and can be found in a multitude of AOPs. Thus, AOPs can be assembled into complex Adverse Outcome Networks (AONs) to better evaluate and identify the risks caused by exposure to certain stressors on health and the environment and help reduce the use of animal methods.

Given the enormous amount of existing knowledge in the scientific literature and other data sources, the identification of relevant biological information to build AOP is a complex and time-consuming task, especially in deciphering dispersed data. Thus, to assist in the AOP development process, the AOP-helpFinder tool has been developed [2]. AOP-helpFinder is an innovative tool, developed under python, combining graph theory and text mining that automatically explores scientific abstracts from the PubMed database.
From a list of stressors (e.g., "bisphenol", "ionizing radiation") and a second list of key biological events (e.g., "oxidative stress", "DNA breaks") provided by the user, AOP-helpFinder will be able to identify and extract, in a systematic manner, all published existing associations between at least one stressor and one key events from the lists. For example, the tool has already been successfully applied for the identification of links between different substances (bisphenol S, bisphenol F and pesticides) and KEs [2, 3, 4]; as well as for the development of an AOP on microcephaly induced by ionizing radiation (https://aopwiki.org/aops/441). Recently, a web-server (available at: http://aop-helpfinder.u-paris-sciences.fr/) has also been developed for a friendly use of the tool [5].
In the current version, AOP-helpFinder only allows the search for links between stressors and KEs/AOs. An updated version is under development to also offer a search for links between pairs of biological events. The ultimate goal is to provide a tool to assist experts at all levels of AOP development but also to provide automated creation of predictive AOPs and AOP networks, which would greatly support the development of systems toxicology.

References

[1] G. T. Ankley *et al.*, "Adverse outcome pathways: a conceptual framework to support ecotoxicology research and risk assessment," *Environ. Toxicol. Chem.*, vol. 29, no. 3, pp. 730–741, Mar. 2010, doi: 10.1002/etc.34.

[2] J.-C. Carvaillo, R. Barouki, X. Coumoul, and K. Audouze, "Linking Bisphenol S to Adverse Outcome Pathways Using a Combined Text Mining and Systems Biology Approach," *Environ. Health Perspect.*, vol. 127, no. 4, p. 47005, Apr. 2019, doi: 10.1289/EHP4200.

[3] M. Rugard, X. Coumoul, J.-C. Carvaillo, R. Barouki, and K. Audouze, "Deciphering Adverse Outcome Pathway Network Linked to Bisphenol F Using Text Mining and Systems Toxicology Approaches," *Toxicol. Sci. Off. J. Soc. Toxicol.*, vol. 173, no. 1, pp. 32–40, Jan. 2020, doi: 10.1093/toxsci/kfz214.

[4] F. Jornod *et al.*, "AOP4EUpest: mapping of pesticides in adverse outcome pathways using a text mining tool," *Bioinformatics*, vol. 36, no. 15, pp. 4379–4381, Aug. 2020, doi: 10.1093/bioinformatics/btaa545.

[5] F. Jornod, T. Jaylet, L. Blaha, D. Sarigiannis, L. Tamisier, and K. Audouze, "AOP-helpFinder webserver: a tool for comprehensive analysis of the literature to support adverse outcome pathways development," *Bioinformatics*, vol. 38, no. 4, pp. 1173–1175, Feb. 2022, doi: 10.1093/bioinformatics/btab750.

Feature selection in longitudinal microbiome data via the analysis of random projections

Antonella Gieco[1], Sébastien Leuillet[2] and Diego Tomassi[2]
[1] Universidad Nacional del Litoral, Santiago del Estero 2829, 3000 Santa Fe, Argentine
[2] Biofortis SAS, 3 route de la Chatterie, 44800 Saint-Herblain, France

Corresponding author: diego.tomassi@biofortis.fr

1 Introduction

There is increasing interest in studying the connection between some diseases and the gut microbiota, as well as to identify what specific microbes play a fundamental role in that. Formally, this can be cast as a variable selection problem. Despite standard in statistics, it poses a big challenge to available methods to deal simultaneously with the intrinsic complexity of the data and the experiment design. Briefly, methods that can accomodate specific designs and longitudinal data tend to be univariate, while multivariate methods often do not take full advantage of the repeated measurements. In this work we present a two-step approach that combines the main advantages of univariate approaches, while still preserving a multivariate nature via random linear combinations of the original features.

2 Proposed method

Let $\mathbf{X} = (X_1, \ldots, X_T)$ be a sequence of microbiome profiles $X_t \in \mathbb{R}^q$ collected at times $t = 1, \ldots, T$ for the same individual and let $\mathbf{u}_k \in \mathbb{R}^q$ be a fixed vector of norm one. Let $\mathbf{Z}_k = (Z_{1k}, \ldots, Z_{Tk})$, with $Z_{tk} = \langle \mathbf{u}_k, X_t \rangle$, be the sequence of projections of the microbiome profiles onto the direction $\mathbf{u}_k$. Also, let $\mathcal{L}(\mathbf{Z}_k, \mathbf{Y})$ denote a linear mixed model for the outcome $\mathbf{Z}_k$ as a function of the fixed effects in $\mathbf{Y}$, using the subject ID as a random effect. The proposed method consists of the following two steps:

- **Step 1 (Tested random projections):** Let $\tilde{\mathbf{U}} = \{\mathbf{u}_1, \mathbf{u}_2, \ldots, \mathbf{u}_p\}$ be a set of p randomly generated directions of projection, with large enough p. Get the matrix $\mathbf{U}$ of column vectors in $\tilde{\mathbf{U}}$ so that $\mathbf{u}_k$ is a column of $\mathbf{U}$ iff $\mathcal{L}(\mathbf{Z}_k, \mathbf{Y})$ gets a significant p-value for the intended comparison on $\mathbf{Z}_k$. That is, $\mathbf{U}$ is built only from the projective directions that are good enough to discriminate between the tested groups in $\mathbf{Y}$. No correction for simultaneous testing on the p directions is included at this step.
- **Step 2 (Sparsity-inducing matrix factorization):** Find a reduced-rank factorization of $\mathbf{U}$ as $\mathbf{U} = \mathbf{AB}$, inducing some full rows of matrix $\mathbf{A}$ to go to zero. Then, if $\mathcal{I}$ indexes the set of rows of $\mathbf{A}$ that are not null, it determines the set of relevant features as $X_{t(1)}, \ldots, X_{t(\ell)}$ for indices $(1), \ldots, (\ell)$ in $\mathcal{I}$ and for every time-point t.

Step 1 evaluates a set of random candidate directions to project the data and retains only those who are discriminant. Since each projection is scalar, this can be achieved using standard tools that easily accomodate repeated measurements. For large enough p, matrix $\mathbf{U}$ spans a subspace that retains key information about $\mathbf{Y}$ available on $\mathbf{X}$, while avoiding multivariate estimation. Since each $\mathbf{u}_k$ combines information from all the variables, the method preserves a multivariate nature. Moreover, since $\mathbf{U}$ can be made rank-defficient, so that it has redundant information to characterize such a subspace, the factorization in **Step 2** finds a minimal spanning basis $\mathbf{A}$ of the discriminant subspace while inducing variable selection via structured sparsity. Provided the dimension of such minimal subspace is known (i.e. by testing on the SVD of $\mathbf{U}$), the optimization problem can be cast as a group-lasso problem.

3 Results

Obtained results from experiments with GAN-based simulated microbiome data [1] show that the proposed method is competitive in terms of selection accuracy (false discoveries, false negatives), while not requiring complex estimation of multivariate models but software tools commonly available.

References

[1] R. Rong, S. Jiang, L. Xu, G. Xiao, Y. Xie, D. J. Liu, Q. Li, and X. Zhan. MB-GAN: Microbiome Simulation via Generative Adversarial Network. *GigaScience*, 10(2), 02 2021. giab005.

Axonal Delay Learning: from biology to computational neuroscience

Amélie Gruel[1] and Jean Martinet[1]
[1]Université Côte d'Azur, CNRS, I3S, France

Corresponding author: amelie.gruel@univ-cotedazur.fr

Computational neuroscience aims to model the biological reality of brain behaviour using computer science tools. In particular, one branch of this research field has specialised towards bio-inspired artificial intelligence, which comprises Spiking Neural Network (SNN). A spiking neuron mimics the dynamics of biological neuronal circuits by receiving, processing and sending information in the form of spike trains [1]. A SNN is constructed using populations of neurons linked together with connections, with respect to a certain architecture and certain rules allowing it to learn a behaviour. These rules often concern the evolution of the synaptic weight between two pre- and post-synaptic neurons, i.e. the amount of neurotransmitters released in the synapse following the pre-synaptic neuron's excitation.

Delay learning is another of those learning rules which, instead of changing the synaptic weight, changes the delay of the electrical spike journey in the pre-synaptic neuron's axon [2].

First biological observations

The first evidence of any neuronal delay in the information propagation within the animal brain came from the interaural time difference, allowing for the azimutal localization of sound by barn owls. According to Gerstner in 1996, there is a true paradox in auditory neural systems since "neural networks encode behaviourally relevant signals in the range of a few μs with neurons that are at least one order of magnitude slower" [3]. Various biological experiments have thus revealed the existence of a biological axonal delay precisely adjusted according to variations of parameters in the brainstem.

The importance of myelination

Over and over, myelin has been identified as one of the parameters mentioned above. Indeed, this multilaminar coating formed by the glial cells in the Vertebrates' nervous system facilitates both the neural circuit function and the behavioural performance [4]. Experiments on mammals show that myelination is directly related to learning and memory consolidation, both at an early age and in older animals, due to its involvement in coupling the activity of distant neuron populations.

Some computational approaches

Delay learning is a striking example of a computer science concept effectively reinforced by ongoing neuroscience work. Indeed, more and more SNN models are being developed with an impetus to learn by updating delays, not just synaptic weights. One convincing instance among many is given by [1], which combines delay adaptation and polychronization for reservoir computing; or more recently by [5], which proposes an STDP extended to the delay learning repeating spatio-temporal patterns. We likewise aim to implement such a delay learning applied to motion detection, notably by drawing on neurological knowledge in order to approach biological efficiency.

Acknowledgements

This work was supported by the European Union's ERA-NET CHIST-ERA 2018 research and innovation programme under grant agreement ANR-19-CHR3-0008.

References

[1] H. Paugam-Moisy et al. Delay learning and polychronization for reservoir computing. *Neurocomp.*, 71, 2008.
[2] H. Hüning et al. Synaptic delay learning in pulse-coupled neurons. *Neural Computation*, 10, 1998.
[3] W. Gerstner et al. A neuronal learning rule for sub-millisecond temporal coding. *Nature*, 383, 1996.
[4] R. Fields. A new mechanism of nervous system plasticity: activity-dependent myelination. *Nature Reviews Neuroscience*, 16, 2015.
[5] A. Nadafian and M. Ganjtabesh. Bio-plausible unsupervised delay learning for extracting temporal features in spiking neural networks. *arXiv*, 2020.

Machine learning analysis on transcriptomic data reveals novel target genes of the WNT/β-catenin pathway in colorectal cancer

Cemre Kefeli[1] and Andrés Aravena[1]
Molecular Biology and Genetics Department, Istanbul University, Istanbul, Turkey

Corresponding author: andres.aravena@istanbul.edu.tr

1 Background

The Wnt signalling pathway is a driving force of proliferation and differentiation [1]. Aberrant behaviour in this pathway may lead to several types of cancers and Alzheimer's disease [2,3]. To complete our understanding of this pathway we look for its target genes, which may be tissue-specific. We found 93 target genes identified experimentally in the colorectal cancer context, but it is probable that the real number is higher. We aim to identify novel target genes of the Wnt/β-catenin signalling pathway using a one-class machine learning approach, expanding the method presented in [4]. The challenge here is the lack of well-defined negative examples.

2 Materials and methods

We analyzed several publicly available transcriptome experiments and used the differential gene expression profiles to represent each gene. We used the experimentally validated target genes as "positive" examples in training. We took a random sample from the rest of the genes as "negative" examples to train a CART classifier. This process was repeated 1000 times with independent sampling. Each trained classifier was then used to assign a "positive" or "negative" label for each gene. The number of times each gene is classified as "positive" is a score that can be tested for significance using the Fisher method. The classification depends on two hyperparameters, sample size and misclassification cost, which we tuned to minimize over-fitting using a suitable negative control. Thus, we found a set of putative target genes having an expression pattern very similar to known target genes.

3 Results

The pool of trained classifiers predicted 144 putative novel target genes. Some of the highest scoring genes are PTCH1, GLI3 and SOX4. The first two predictions, PTCH1 and GLI3, are important components of the Hedgehog Signalling. This suggests a possible interplay between Wnt and Hedgehog signalling in colorectal cancer. In parallel to our study, experimental researchers have shown that PTCH1 is indeed a colon specific Wnt target [5].

4 Conclusion

We present a bioinformatic method to find putative target genes of the canonical Wnt signalling pathway, and eventually in other pathways, based only on gene expression data and a set of experimentally validated targets. This method narrows the set of genes to validate experimentally. Moreover, some of our predictions have already been validated by other studies.

References

[1] Catriona Y. Logan and Roel Nusse. The Wnt signaling pathway in development and disease. *Annual Review of Cell and Developmental Biology*, 20(1):781–810, nov 2004.

[2] Hans Clevers. Wnt/β-catenin signaling in development and disease. *Cell*, 127(3):469–480, nov 2006.

[3] Giancarlo V. De Ferrari and Nibaldo C. Inestrosa. Wnt signaling function in Alzheimer's disease. *Brain Research Reviews*, 33(1):1–12, aug 2000.

[4] Christian Hödar, Rodrigo Assar, Marcela Colombres, Andrés Aravena, Leonardo Pavez, Mauricio González, Servet Martínez, Nibaldo C Inestrosa, and Alejandro Maass. Genome-wide identification of new wnt/β-catenin target genes in the human genome using CART method. *BMC Genomics*, 11(1):348, 2010.

[5] Kim Elisabeth Boonekamp, Inha Heo, Benedetta Artegiani, Priyanca Asra, Gijs van Son, Joep de Ligt, and Hans Clevers. Identification of novel human Wnt target genes using adult endodermal tissue-derived organoids. *Developmental Biology*, 474:37–47, jun 2021.

SciGeneX: an unsupervised method to naturally discover cell types or cell states based on patterns of co-expressed genes in single-cell RNA-sequencing data

Julie BAVAIS[12], Lionel SPINELLI[12] and Denis PUTHIER[1]

[1] Theories and Approaches of Genomic Complexity (TAGC), Aix-Marseille University, INSERM, Turing Centre for Living Systems, 163 avenue de Luminy, 13009, Marseille, France

[2] Centre d'Immunologie de Marseille Luminy (CIML), Aix-Marseille University, INSERM, CNRS, Turing Centre for Living Systems, 163 avenue de Luminy, 13009, Marseille, France

Corresponding Author: bavais@ciml.univ-mrs.fr

Single-cell RNA sequencing revolutionizes transcriptomic studies providing gene expression level at single-cell resolution [1]. The classical pipeline used to discover cell populations consists of several consecutive steps, namely feature selection, dimension reduction and cell clustering [2]. These steps are widely used in the world of single-cell RNA-sequencing, however, three major problems remain to be improved. First, feature selection methods do not lead to a common consensus and most of them provide variable results [3]. Second, the identification of marker genes leads to double dipping by creating a high type I error rate caused by the prior identification of cell groups [4]. Finally, all the steps of the classical pipeline depend on a number of parameters which, depending on their values, will generate a large variability of results.

To overcome these problems, we developed SciGeneX (for Single-cell informative Gene eXplorer). An unsupervised method offering an alternative approach that provides an initial insight into the pattern of co-expressed genes across cells. SciGeneX automatically filters co-expressed genes across the set of cells using a density-based-filtering algorithm and clusters them into gene patterns of expression using the Markov Cluster Algorithm [5]. Combinations of these patterns spontaneously highlight biologically relevant cell populations associated with cell types or states as well as the genes specifically expressed in these populations. Thus, SciGeneX perform feature selection and identification of co-expressed gene patterns and provide an alternative approach for cell clustering based on these patterns, avoiding the main drawbacks of the currently used algorithms.

Acknowledgements

SciGeneX is freely available as a R package on github at https://github.com/dputhier/scigenex for Linux users and is compatible with the classical seurat workflow.

References

[1] Kolodziejczyk, A.A., Kim, J.K., Svensson, V., Marioni, J.C., and Teichmann, S.A., The technology and biology of single-cell RNA sequencing. Mol. Cell 58, 610–620, 2015.

[2] Luecken MD and Theis FJ. Current best practices in single-cell RNA-seq analysis: a tutorial. Mol Syst iol. 2019 Jun 19;15(6):e8746.

[3] Lähnemann, D., Köster, J., Szczurek, E. et al. Eleven grand challenges in single-cell data science. Genome Biol 21, 31, 2020.

[4] Gao, L. L., Bien, J., and Witten, D., Selective Inference for Hierarchical Clustering, ArXiv, 2020.

[5] Stijn van Dongen, Graph Clustering by Flow Simulation. PhD thesis, University of Utrecht, 2000.

Bioinformatics integration of regulatory regions and variants in immune cells

Marie MICHEL[1], Aitor GONZALEZ[1] and Badih GHATTAS[2]

[1] TAGC, 163 avenue de Luminy, 13006, Marseille, France

[2] I2M, 163 avenue de Luminy, 13006, Marseille, France

Corresponding Author: marie.michel@univ-amu.fr

The immune system is capable of detecting and killing various pathogenic agents such as viruses and bacteria but also cells, including cancer cells. In some cases, cancer cells can escape the immune response. Understanding the regulation of the immune system cells is important to know how a patient could react to cancer and cancer treatments. We are interested in the regulation of the immune system, that is the non-coding variants involved in the immune regulation.

In order to do that the first objective was to collect data in the different immune cell types: first, we collected ATAC-seq data [1] which allowed us to study the regulatory regions which are known to contain regulatory variants. ATAC seq is a technique that aims to localize open chromatin regions that are known to be enriched in regulatory elements. For that, a hyperactive transposase inserts sequencing adaptors to the DNA. The tagged DNA fragments are then purified, PCR-amplified, and sequenced using next-generation sequencing. Sequencing reads can then be used to infer regions of increased accessibility. The number of reads for a region correlates with how open that chromatin is, at single nucleotide resolution.

Then the same thing has been done with the eQTL [2] which allows the analysis of the regulatory variants directly. eQTL stand for expression quantitative trait loci and aim to localize loci that have an influence in the expression of genes. The idea was to check whether the different immune cell types are separated in order to check the coherence of the data.

We saw with ATAC-seq that we had a good separation between stimulated and unstimulated conditions but also a good separation between the main cell types whether it is on the basis of normalized peaks or motifs. Regarding eQTL we saw that the separation between stimulated and unstimulated conditions was not clear but well by cell types.

After this we wanted to know whether eQTLs were found in ATAC-seq peaks for similar cell types. For that we used Chromvar [3]. ChromVar computes raw deviation in accessibility, then the raw deviations for background peaks and finally bias corrected deviation and z-score. A high score means an enrichment of eQTLs of the dataset in the ATAC-seq sample. We can see enrichment of eQTL in the ATAC-seq region of the same cell type.

Next, I use deep learning based on ATAC-seq and eQTL DNA sequences as well as the ATAC seq count matrix in order to predict variant prediction. The results of de deep learning show good results.

My next objectives are to collect new data in immune cell types such as CHIP-seq, DNase-seq, histones marks and so on. Finally, we want to link everything to cancer by integrating data from GWAS, TCGA or PCAWG.

References

1. Calderon, D., et al. Landscape of stimulation-responsive chromatin across diverse human immune cells. *Nat Genet*, (51):1494–1505, 2019.
2. Kerimov, N., Hayhurst, J.D., Peikova, K. et al. A compendium of uniformly processed human gene expression and splicing quantitative trait loci. *Nat Genet,* (53):1290–1299, 2021.
3. Schep, A., Wu, B., Buenrostro, J. et al. chromVAR: inferring transcription-factor-associated accessibility from single-cell epigenomic data. Nat Methods, (14):975–978, 2017.

Bulk RNA-Seq deconvolution for the study of hemorrhagic fever

Emeline Perthame[1], Hélène Lopez-Maestre[1] and Natalia Pietrosemoli[1]
Hub of Bioinformatics and Biostatistics, Institut Pasteur, 25-28 rue du Docteur Roux, Paris, France

Corresponding author: emeline.perthame@pasteur.fr

1 Abstract

We are interested in studying the kinetics of Lassa fever infection in Cynomolgus monkeys [1]. Lassa virus produces a viral hemorrhagic fever that has been quickly expanding in the African continent and its pathogenesis and the spectrum of severity is not fully understood.

Here, we compare the virus invasion into several tissues to describe pathogenesis of the disease and host response. By inferring cell type proportions for the different viral strains using deconvolution methods, we can provide keys to understanding the host response. The evolution of the proportions of activated, resident or infiltrated immune cells allows for refined characterisation of the nature of immune activation and tissue inflammation and possibly of the cell populations destroyed by the infection. We apply our methods to a transcriptomic dataset (RNAseq) representing the evolution of Lassa infection at 3 time points for healthy and infected monkeys. The comparison of a lethal and a non-lethal strain of the virus allows the identification of markers both of early infection and severity. Currently, there are more than 50 cell type deconvolution methods available in the literature, mostly implemented on variations of the linear model [2]. We selected a handful of methods according to their "feasibility" such as, the most used in the community, R implementation into a well maintained and easy to install package, documentation on the method, and running time, as assessed in recent reviews of bulk RNA-Seq deconvolution methods [3], [2]. Our analysis includes methods such as CIBERSORT [4], one of the most used methods in the community, as well as the Ordinary Least Square which represents a straight-forward implementation to estimate the regression coefficients reflecting cell types abundances.

Most deconvolution methods require two inputs: the count matrix representing the gene expression values for each sample M, and a matrix of signatures S, required to define the gene expression levels corresponding to each cell type. The S matrix needs to be specific to the tissue or organ under study. In order to refine the signature definition, we designed a two-level procedure: first, inferring the cell types proportions with a signature specific to the tissues under study, to identify the main cell types, and second, computing the remaining proportions with three publicly available signatures for immune cells. We show that the construction of S has a substantial impact on the results, as missing cell types can alter the proportions of the inferred cells and thus produce misleading results.

Last, we discuss some key elements to biologically interpret the results provided by this analysis. For example, the application of CIBERSORT to liver cells using one of our combined signature shows that the proportion of activated NK cells, CD4 T cells and naive B cells increases sharply in the tissue during the course of the disease, which shows the high level of inflammation in the liver. This observation is confirmed using histology.

Acknowledgements

We acknowledge the Biology of viral emerging infections unit (Institut Pasteur, Lyon) directed by Sylvain Baize for providing their data and biological expertise to interpret the results.

References

[1] N. Baillet and co authors. Systemic viral spreading and defective host responses are associated with fatal Lassa fever in macaques. *Communications Biology*, 4(1).

[2] H. Jin and Z. Liu. A benchmark for RNA-seq deconvolution analysis under dynamic testing environments. *Genome Biology*, 22(1):102, December 2021.

[3] F. Avila Cobos and co authors. Benchmarking of cell type deconvolution pipelines for transcriptomics data. *Nature Communications*, 11(1).

[4] A.M. Newman and co authors. Robust enumeration of cell subsets from tissue expression profiles. *Nature Methods*, 12(5).

DetecTree, from freehand drawing to digital patient care in genomic medicine

Rafik Mankour[1], Jiri Ruzicka[1], Stella Wolff[1], Candice Hermant[1], Denis Bertrand[1], Nicolas Duforet-Frebourg[1] and Kévin Yauy[1,2]

[1] SeqOne Genomics, 22 Rue Durand, 34000, Montpellier, France

[2] Institute of Advanced Biosciences, Centre de recherche UGA, Inserm U 1209, CNRS UMR 5309, Grenoble, France.

Corresponding Author: rafik.mankour@seqone.com

Keywords Genomic medicine, family tree, genogram, computer vision, Detectron2

1 Background/Objectives

Synthesizing the hereditary information of each family case, family trees or genograms are a central tool for patient management and variant interpretation in genomic medicine. However, as it is currently produced on paper during consultations, this clinical information is not exploitable in bioinformatic analysis pipelines. Despite the availability of digital drawing tools, there is unfortunately no adoption in the community yet.

2 Methods

We develop DetecTree, an artificial intelligence system able to automatically digitize family trees, including the detection of the number and position of the family members, their attributes and their relationships. In May 2022, we collected and annotated manually 38 digitally made genograms and 47 photographed images of anonymous family trees drawn by health professionals as in the usual practice [1]. We trained a deep learning model based on the Detectron2 framework [2] that takes a family tree picture as an input and produces a digitalized family tree as an output compliant with standard formats.

3 Results

We achieved promising results, managing to correctly register 95% of the individuals of our evaluation batch in the right generation and order. We can confidently identify the sex and disease status attributes with an AP50 of around 95%, but still get limited performance for relationship detection, and deceased and proband status detection because of the low representation of these cases in our initial samples. Using these predictions, DetecTree attributes for each family member its position in their respective generation and detected parent-offspring relationships.

4 Conclusion

DetecTree is a deep learning system for genogram digitalization. It yields exploitable outputs in automated digital DNA analysis and thus could improve the diagnostic management of patients. Our solution passed the initial testing phase and we are scaling it to manage more diverse data inputs by retraining and optimizing our models on the real-life collected data, with a more balanced representation of all characteristics to identify deceased status and proband position.

References

[1] RL Bennett, KA Steinhaus, SB Uhrich, et al. Recommendations for standardized human pedigree nomenclature. *American journal of human genetics* vol. 56,3 (1995): *745-52.*

[2] Y. Wu, A. Kirillov, F. Massa, W.-Y. Lo, and R. Girshick. Detectron2. https://github.com/facebookresearch/detectron2, 2019.

Wasserstein regularisation for multidataset PCA

Stéphane Béreux[1−2], Magali Berland[1], Sébastien Fromentin[1] and Mahendra Mariadassou[2]
[1] INRAE - MétaGénoPolis, Domaine de Vilvert, 78352, Jouy-en-Josas, France
[2] INRAE - MaIAGE, Domaine de Vilvert, 78352, Jouy-en-Josas, France

Corresponding author: stephane.bereux@inrae.fr

1 Abstract

Health data are often very high-dimensional (especially omics and metaomics) and difficult to collect in large quantities, due to resource constraints and/or disease rarity.
Classically, n refers to the number of samples (*e.g.* patients in a cohort) while p refers to the number of features (*e.g.* number of gene expressions for transcriptomic data, the number of species for microbiome data, etc). To address the high-dimensional regime ($p \ll n$) where many standard approaches fail, several steps are usually taken:

1. The first is to get more data (i.e. increase n). For example, in the context of medical data, it is common to have to gather several datasets, from different cohorts, to increase the amount of data, especially when the size of each dataset is limited.
2. The second is to reduce the size of the data (i.e. decrease p). Principal Component Analysis (PCA) remains a very popular first approach to dimension reduction, due to its good performance and ease of use.
3. The third is to use regularisation techniques (such as LASSO) to perform feature selection (i.e. choose only some features).

The focus here is on the case where there are several datasets measuring the same features on several cohorts of samples (multi-block approch). A naive way of integrating the data consists in concatenating these datasets, in line with point 1), in order to increase n. They are likely mutually informative, even though they probably each have their own specificity. One could then apply a dimension reduction technique, such as PCA, to decrease p as in 2).
However, this simple approach leaves us vulnerable to batch effects which might bias the global analyses. It is there therefore recommended to both pool the different datasets but still correct for the peculiarity of each dataset. We propose here an approach based on Wasserstein regularisation and Optimal Transport [1] to allow a penalised variability between the loadings obtained by an independent PCA on each of the individual datasets, which we call Multi-Wasserstein PCA (MWPCA).
More formally, given a collection $\mathbb{X}$ of D datasets, $\nu_S, \nu_L, \mu \in \mathbb{R}_*^+$, and a rank $k \in \mathbb{N}^*$, the MWPCA consists in finding a collection $(\mathbb{S}, \mathbb{L})$ of D pairs $(S^d \in \mathbb{R}^{n \times k}, L^d \in \mathbb{R}^{k \times p})$, which minimizes the loss:

$$\frac{1}{D}\sum_{d=1}^{D}\frac{1}{n_d}\left(\underbrace{||X^d - S^d L^d||_F^2}_{\textbf{Reconstruction error}} + \underbrace{\nu_S \sum_{l=1}^{k} ||S_{.l}^d||_2}_{\textbf{Norm regularisation}} + \underbrace{\nu_L\, ||L^d||_1}_{\substack{\textbf{Norm regularisation} \\ \textbf{+ Sparsity}}}\right) + \underbrace{\mu\, \Omega(L^1, ..., L^D)}_{\substack{\textbf{Wasserstein regularisation} \\ \implies \textbf{Loadings consistency}}}$$

This model is computed using an alternating minimisation algorithm [2], combined with a penalty based on Sinkhorn's algorithm [3] to ensure inter-dataset consistency of the loading matrices.

We demonstrate our methodology on microbiome data, for which it is known that there are country-specific compositional differences. We predict colorectal cancer (CRC), based on the patient's microbiome, applying MWPCA to nine datasets from studies conducted in various countries.

References

[1] Hicham Janati, Marco Cuturi, and Alexandre Gramfort. Wasserstein regularization for sparse multi-task regression. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 1407–1416. PMLR, 2019.

[2] Madeleine Udell. *Generalized low rank models*. Stanford University, 2015.

[3] Marco Cuturi. Sinkhorn distances: Lightspeed computation of optimal transport. *Advances in neural information processing systems*, 26:2292–2300, 2013.

Predicting the tissue of origin from circulating DNA fragments: Biological lessons learned from a comprehensive analysis of genetic, functional and computational features to increase the accuracy of a statistical mode.

Elyas MOUHOU[1], Josselin NOIREL[1] and Charlotte PROUDHON[2]

[1] CNAM, Laboratoire GBCM, 2 rue Conté, 75003, Paris, France

[2] Institut Curie, INSERM U934/CNRS UMR3215, PSL Research University, Paris, France

Corresponding Author: josselin.noirel@lecnam.net

Abstract

Several studies have made it possible to envision a translational application of plasma sequencing in cancer diagnosis and monitoring[1]. However, for early cancer detection, the stoichiometry of fragments tumour cell DNA (ctDNA) among the cell-free DNA (cfDNA) remains a formidable challenge to overcome.

Dying cells, even in healthy individuals, release a fraction of the digested fragments of their genetic material into the blood stream. Interestingly, the position of the nucleosomes remains imprinted onto the circulating DNA fragments so that those fragments can be sequenced and statistical models can be trained to recognise the tissue those fragments originate from.

This information, if made sensitive enough, could be a useful medical device to carry out so-called 'liquid biopsies', allowing clinicians to diagnose at an early stage or to precisely monitor a number of diseases, including cancer. In this study, we comprehensively study genetic and computational features of circulating DNA fragments in the vicinity of transcriptional start sites that increases the sensitivity of statistical models trained on the sequences of circulating DNA from healthy individuals.

In this study, we set about comprehensively appraising the predictive value of genetic, functional as well as computational features of cfDNA around transcription start sites (TSSs) in models trained to identify the (lympho-myeloid) tissue(s) of origin of cfDNA in healthy samples from public datasets.

Our study also considers a variety of selection approaches to derive sets of tissue-specific genes. The features of interest are: mapped fragment density around each TSS, inclusion of long non-coding RNA (lncRNA) genes, inclusion of the first nucleosome position after a TTS. The inclusion or not of each feature together with the selection approach was evaluated combinatorially in an attempt to increase the performance of those models but also to derive biological insight into the gene expression landscape seen from a tissue perspective.

Our results show that:

1. lncRNAs are more tissue-specific than coding genes;
2. The first nucleosome position after a TTS doesn't convey a substantial amount of predictive information to identify the tissue of origin;
3. The comparison of highly specific genes vs non-specific genes are the most useful predictive comparison.

While the main approach in the field tend to rely more and more on deeplearning[1,2], leading to blackbox model, our approach offer to improve the biological knowledge of the biomarker that is the cfDNA.

References

1. Stephen Christiano et al.. Genome-wide cell-free DNA fragmentation in patients with cancer, *Nature* volume 570, pages385–389 (2019).
2. Jiaqi Li et al., DISMIR: Deep learning-based noninvasive cancer detection by integrating DNA sequence and methylation information of individual cell-free DNA reads, *Briefings in Bioinformatics*, Volume 22, Issue 6, November 2021.

Computational study of chemical-induced liver injury using high-content imaging phenotypes

Vanille LEJAL[1], David ROUQUIÉ[2] and Olivier TABOUREAU[1]

[1] Université Paris Cité, INSERM U1133, CNRS UMR 8251, 75006, Paris, France

[2] Bayer SAS, Bayer Crop Science, 355 rue Dostoïevski, CS 90153, 06906, Valbonne, Sophia-Antipolis, France

Corresponding Author: vanille.lejal@etu.u-paris.fr

Toxicological studies are designed to characterize the adverse effects of chemicals according to the dose and the duration of exposure to ensure their safety for humans. These past years, guided by progress in data science and the objective to reduce animal testing, new *in silico* and *in vitro* methods appeared as a supplement to *in vivo* studies. In this context, the Broad Institute of the MIT developed the Cell Painting [1], a technique of image-based cell profiling, to study morphological changes induced by chemicals in cells cultivated *in vitro* and to identify potential toxic risks.

In this study, we focused on how Cell Painting could help to predict drug-induced liver injury (DILI). Overall, 537 chemicals from the Cell Painting data were gathered and annotated using two datasets of hepatotoxicity findings in humans and rodents called DILIrank [2] and eTox [3]. Based on an ensemble of 1779 morphological features, a T-test was performed to identify features that would be more likely to discriminate between DILI and non-DILI chemicals. Then machine learning methods were computed to assess the hepatotoxicity risk induced by the chemical.

Combined with 2 times 5-fold cross-validation and grid-search hyperparameter tuning, the ElasticNet regression model gave encouraging predictive performances with balanced accuracies (BA) higher than 0.6 for test sets. Morphological features related to cells' appearance, especially roughness and smoothness, seem to contribute to the performance of the model. Our results show that the application of machine learning tools on phenotypic screening data could be used for chemical risk assessment in hepatic failures.

In the near future, morphological profiles induced in cells by chemicals could be a new lead to improve the identification of chemical-phenotype relationships of interest not only in toxicology but also in pharmacology.

References

1. Mark-Anthony Bray, Shantanu Singh, Han Han, Chadwick T Davis, Blake Borgeson, Cathy Hartland, Maria Kost-Alimova, Sigrun M Gustafsdottir, Christopher C Gibson and Anne E Carpenter. Cell Painting, a high-content image-based assay for morphological profiling using multiplexed fluorescent dyes. *Nature Protocols*, (11):1757-1774, 2016.
2. Minjun Chen, Ayako Suzuki, Shraddha Thakkar, Ke Yu, Chuchu Hu and Weida Tong. DILIrank: the largest reference drug list ranked by the risk for developing drug-induced liver injury in humans. *Drug Discovery Today*, (4):648-653, 2016.
3. Denis Mulliner, Friedemann Schmidt, Manuela Stolte, Hans-Peter Spirkl, Andreas Czich, and Alexander Amberg. Computational Models for Human and Animal Hepatotoxicity with a Global Application Scope. *Chemical Research in Toxicology*, 29(5):757-767, 2016.

Predicting adverse drug reactions on organs using sequential neural network models.

Bryan DAFNIET[1], Olivier TABOUREAU[1]

[1]Université de Paris, BFA, UMR 8251, CNRS, ERL U1133, Inserm, F-75013 Paris, France.

Corresponding Author: bryan.dafniet@u-paris.fr

Adverse drug reactions (ADRs) from drugs are a major issue in modern healthcare causing longer hospital stays and higher mortality rates with costs estimated at around 30 billion dollars in the US only [1].

These adverse effects are usually linked to overdosage, drug-drug interactions or the effect of polymorphism on specific therapeutic targets [2]. This study will focus on the latter to develop a predictive model able to assess if an ADR can be caused by a genetic mutation present in a drug target. Mutations used in this study were missense SNPs or SNVs.

Firstly, we used a system pharmacology network integrating data from multiple sources like DrugBank, DrugCentral or dbSNP to compile the information on molecule-target-mutation and adverse effects [3]. Then, we built sequential neural network models to predict the possibility of a compound being associated with ADRs from these extracted data. For simplicity, we grouped the AE into organs using the system organ class (SOC). Three distinct approaches were considered, i) a model using all the SOCs at the same time ii) a distinct model for each SOC, and finally, iii) a model including structural information on the compounds.

The models developed for each SOC were promising with a balanced accuracy of around 0,6 on a training set and 0,57 on the test for the SOC "Respiratory, thoracic and mediastinal disorders". Even more, when structural information is added results are even improved with a notable 0,69 on a training set and 0,63 on the test for the "Metabolism and nutrition disorders" SOC.

With the increasing number of large studies investigating the impacts of pharmacogenomic i.e., the role of the genome in drug responses in patients, such computational implementation could be of interest in the development of personalized medicine where ADRs could be avoided based on the patient genome.

References

1. Sultana J et al. Clinical and economic burden of adverse drug reactions. *J Pharmacol Pharmacother* (4): S73-7, 2013.
2. Yan M et al. Association between gene polymorphism and adverse effects in cancer patients receiving docetaxel treatment: a meta-analysis *Cancer Chemother Pharmacol* (89):173-181. 2022.
3. Dafniet B, et al. Drug-target-ADR Network and Possible Implications of Structural Variants in Adverse Events. *Mol Inform.* (39):1-12, 2020.

Exploring cell morphological profile information for the de-risking of small molecules.

Fabrice CAMILLERI[12], Jean-Paul COMET[1] and David ROUQUIE[2]

[1] I3S UMR 6070 du CNRS, Université Côte d'Azur, Bâtiment Algorithme-Euclide-B, 2000 Route des Lucioles, B.P. 121, 06903 Sophia Antipolis, France

[2] Bayer SAS, Crop Science Division, 355, rue Fedor Dostoïevski, 06906 Sophia-Antipolis Cedex, France

Corresponding Author: camilleri@i3s.unice.fr

Abstract

The phytopharmaceutical and pharmaceutical industries aim to design chemical products that are active on desired targets while limiting the number of off-targets, to make the products safe for humans and the environment in the conditions of use.

The development of such small molecules is a long and costly process. It takes typically about 12 years to deliver a new molecule on the market, costing several hundred million euros. At any moment, a compound can be stopped due to its suboptimal safety profile.

It is then key to evaluate the safety of chemicals as early as possible in the R&D process.

Due to animal welfare concerns, in vivo experiments must be reduced then stopped by 2035[1].

New approach methods (NAM) are being developed [2]. Those new approaches are ambitioning to shift both hazard identification and risk assessment derived from laboratory animal evaluations towards in silico and in vitro models based evaluations.

Those methods are at their beginning, not yet implemented and not yet legally accepted, but they would allow safety evaluation and risk assessment in a fastest, cheapest way and in a more holistic manner while being more ethical and providing sufficient level of protection of human health.

The Cell Painting in vitro assay is one of the promising NAMs with scalability potential, rich biological information content and, when combined with ADME properties, could provide predictive in vivo adverse outcomes.

We are exploring the information contained in cell painting images for the small molecule safety evaluation and human risk assessment. Morphological profiles will be analyzed in the frame of acute toxicity prediction, for Point of Departure (POD) determination, and for the prediction of other in vitro assay results. U2OS cells are currently being used as a first approach, but we believe that a specific set of cell lines need to be defined, to reflect a maximum of derived toxicological endpoints. Moreover, toxicokinetic models will be also discussed in their applications for reverse dosimetry to extrapolate in vivo doses that would produce POD in rat plasma. Those doses will be compared to known in vivo PODs to evaluate their potential to estimate in vivo POD. Finally, the analysis of links between morphological profiles and toxicological outcomes, would allow us to define unsafe morphological profiles, that could then be used to drive the generation of de novo drug that are safe by design, using generative models such as GANs or Autoencoders.

References

1. David Grimm. U.S. EPA to eliminate all mammal testing by 2035. Science, 2019.
2. U.S. EPA. Strategic Vision for Adopting New Approach Methodologies

Automatic Empirical Segmentation of the Peritumoral Area in Lung Cancer Computed Tomography, Locating the Non-anatomical

Alexis NOLIN-LAPALME[1,2], Kim PHAN[2], Tess BERTHIER[2], Robert AVRAM[1] and Julie HUSSIN[1]

[1] Montreal Heart Institute, 5000 Rue Belanger, H1T 1C8, Montreal, Canada
[2] Imagia, 6650 Rue Saint-Urbain #100, H2S 3G9, Montreal, Canada

Corresponding Author: alexis.nolin-lapalme@umontreal.ca

Lung cancer (LC) remains the principal source of cancer-related mortality worldwide [1]. The cornerstone of LC diagnosis and treatment management is based on computed tomography (CT) which allows a noninvasive internal view of the patient's anatomy. This modality has allowed the use of quantitative imaging approaches in attempt to extract quantifiable information from CT images in attempt to design classifying or predictive models [2]. However, most of these approaches focus on the tumor-bounded region thus wasting other potential hotspots of data present within the lung. One of these candidate regions is the peritumoral area (PA); a non-anatomical zone of lung parenchyma surrounding tumors, associated with disease severity and evolution where cancer cells or immune infiltrate can be detected [3]. However, no studies have been able to propose an accurate empirical segmentation approach thus hindering it's use in feature extraction.

We initially used the RIDER dataset [4]. An ensemble of 32 pairs of LC CT taken fifteen minutes apart. Each segmented tumor mask was dilated with a disk-shape structuring element of radius of 3,6,9,12 mm yielding an associated donut-shaped mask approximating the tumor's PA. Lung masks were adjusted using a U-Network-derived mask used to prevent the sampling of non-lung voxels. Using this mask, features were extracted using the PyRadiomics pipeline [5]. Features within image pairs were then compared using their concordance correlation coefficient (CCC) to identify time-stable features. These were subsequently analyzed within patients by comparing each peritumoral dilation with a randomly sampled healthy area and the tumoral area in each CT from the 1018 cases present in the LIDC dataset [6]. The feature map for the LIDC dataset based on the final features describing the PA were then generated and used to generate the final PA segmentation.

Our results suggest that gray-level co-occurrence matrix maximal correlation coefficient as well as skewness describe the unique appearance of the PA and that those features can be used to generate a PA mask. Moreover, these maps show previously unknown patterns such as the connection of the PA to key metastatic pathways. We will validate the information gain offered by the PA will be compared by comparing tumor malignancy prediction from ResNet18-extracted features of the tumor alone or with the generated PA or the tumor dilation. The tumor and the PA will be considered superior if the accuracy significantly exceeds the other alternatives.

In conclusion, our approach demonstrates for the first time that an empirical segmentation of the PA is possible and that the incorporation of that area in quantitative imaging could further improve algorithmic performance in the context of LC-specific quantitative imaging approach. To validate those results in other cancers, this approach will also be attempted brain as well as liver cancer CT datasets.

References

1. Rebecca Siegel, Kimberly Miller, Hannah E. Fuchs and Ahmedin Jemal. Cancer statistics, 2022, In CA: A Cancer Journal for Clinicians. 7–33, 2022.
2. Huanhuan Li, Long Gao, He Ma, Dooman Arefan, Jiachuan He, Jiaqi Wang and Hu Liu. Radiomics-Based Features for Prediction of Histological Subtypes in Central Lung Cancer. Frontier of Oncology. 11:658887, 2021
3. Zhaofeng Tan, Haibin Xue, Yuli Sun, Chuanlong Zhang, Yonglei Song, and Yuanfu Qi. The Role of Tumor Inflammatory Microenvironment in Lung Cancer. Frontier in Pharmacology. 12:688625, 2021
4. RIDER Research Group. The Reference Image Database to Evaluate Response to therapy in lung cancer (RIDER) project: a resource for the development of change-analysis software. Clinical Pharmacology and Therapeutics, 84(4), 448–456, 2008.
5. Joost van Griethuysen, Andriy Fedorov, Chintan Parmar, Ahmed Hosny, Nicole Aucoin, Vivek Narayan, Regina Beets-Tan, Jean-Christophe Fillion-Robin, Steve Pieper, and Hugo Aerts. Computational Radiomics System to Decode the Radiographic Phenotype. Cancer Research. 77(21): 104–107, 2017.
6. LIDC-IDRI Research Group. The Lung Image Database Consortium (LIDC) and Image Database Resource Initiative (IDRI): a completed reference database of lung nodules on CT scans. Medical Physics, 38(2), 915–931, 2011.

Latent Dirichlet Allocation for Double Clustering (LDA-DC): Discovering patients phenotypes and cell populations within a single Bayesian framework

Elie-Julien El Hachem[1], Nataliya Sokolovska[1] and Hédi Soula[1]
Sorbonne University, INSERM, Nutrition and Obesities: systemic approaches, NutriOmique, 75013, Paris

Corresponding author: elie-julien.el_hachem@sorbonne-universite.fr

1 Introduction

Human disorders have a highly multifactorial nature and depend on genetic, behavioral, socio-economic, and environmental factors. The number of metabolic diseases, cancer, and autoimmune pathologies has increased significantly in recent years, making research in this field a public health priority. In parallel, bioclinical routine datasets have expanded in conjunction with all kind of "omics" data, from both the host and microbiota, as well as metabolomic, proteomic, and cytometry data [1]. All these types of data have some underlying structure on their own, taking values on different scales, with different variability, and are differently distributed. In addition, human patients are an equally important source of variability even among carefully selected cohorts: phenotypic variability (age, gender, previous conditions), dietary habits, bad vs good responders to the treatment, etc. In particular, new types of data have emerged which yield description at the cell level ie cytometry of sc RNA seq. These data add a new layer of structuration that needs to be taken into account.

2 Motivations and Results

From the analytical viewpoint, the single cell data are huge-dimensional matrices produced for each subject. The data dimension, i.e., the number of cells, vary from one individual to another, and note that cell types, as well as the correspondence between the cell populations of the subjects, have to be identified before applying any statistical machine learning method. We refer to the challenge we introduce and consider here as to a *double clustering* problem, where the aim is to simultaneously, purely from observations without any prior knowledge determine cell types, as well as stratify patients in order to study mechanisms of pathologies explained by particular cell subpopulations. We propose a novel approach to stratify cell-based observations within a single probabilistic framework, i.e., to extract meaningful phenotype from both patients and cells simultaneously. Our method is a practical extension of the Latent Dirichlet Allocation [2] and is used to solve the Double Clustering task (LDA-DC).The first step of our framework is the identification of the cell types. Once the cell types are fixed, we can efficiently estimate both probability of a phenotype given a patient and the probability of a cell type given a phenotype. We tested our method on different datasets ranging from simulated patients to whom with AML (acute myeloid leukemia) or Crohn's disease, and were able to identify simultaneously clusters of patients and clusters of cells related to patients' conditions. Furthermore, using a network approach, we were able to stratify patients and identify groups of patients with specific phenotypes.

References

[1] C. Manzoni, D. A. Kia, J. Vandrovcova, J. Hardy, N. W. Wood, P. A. Lewis, and R. Ferrari. Genome, transcriptome and proteome: the rise of omics data and their integration in biomedical sciences. *Briefings in Bioinformatics*, 19(2):286–302, November 2016.

[2] D. Blei, A. Ng, and M. Jordan. Latent Dirichlet allocation. *Journal Of Machine Learning Research*, 3:993–1022, 2003.

Chemomaps: Exploring the chimiodiversity of the living organisms

Solweig Hennechart[1,2], Guillaume Marti[1] and Guillaume Cabanac[2]
[1] Université Paul Sabatier, LRSV, 24 chemin de Borde Rouge, 31326 Castanet-Tolosan, France
[2] Université Paul Sabatier, IRIT, 118 route de Narbonne, 31062 Toulouse, France

Corresponding author: solweig.hennechart@univ-tlse3.fr

The chemodiversity of living organisms corresponds to all natural compounds, and is organized around taxa, trophic environments and ecosystems in a dynamic and a cartography that is still poorly documented. This is obviously part of the problem of biodiversity erosion, thus limiting pharmacological and agri-food knowledge that can be pivotal. The experimental study of this chemodiversity is done using a metabolomic approach, the last element of 'omics' cascade.

Since the year 2000, more than 120 of public or licensed compounds databases have been created [1]. They are more or less specialised in chemical classes such as LipidMaps or in taxa such as HMDB. Their structures and contents are quite different because the objectives are not the same. The Chemomaps project aims to fill these gaps by building a database of all known natural products. It involves integrating and merging data from several heterogeneous sources.

The internal observation of the metabolomics team at the origin of this project is that the data processing during the compound annotation stage is not optimal currently. Indeed, there is no data source associated with a processing chain that performs the annotation process including filters on the biological origin as well as on the structurally close molecules. These criteria should be included to prioritize candidate results, so as to optimize and accelerate the compound annotation. Since 2017, our team has developed initial versions of a database merging several data sources, as well as a processing tool for annotation: MS-CleanR [2]. This work confirmed the interest of this database, but also highlighted the limitations that led to the continuation of this project with a multidisciplinary PhD to rethink the construction of the Pharmakon database to meet new needs. We especially seek to correct the lack of information on the biological origin of compounds. Indeed, this information is essential but missing for two-thirds of the compounds retrieved in Pharmakon2020, the latest version including more than 600,000 compounds.

Based on the observed distribution of chemical classes according to the phylogenetic branches (plant, animal, bacterium, and fungus) [3], this project proposes to develop inference models to predict the biological origin of compounds. We hope to confirm the results obtained on a reduced dataset from Pharmakon2020: our approach yielded a 80% precision when predicting the correct phylogenetic branch for 9,736 compounds described with a single biosource. An experimental validation step of the predictions will be performed on organisms (plants) to confirm or improve the models.

References

[1] Maria Sorokina and Christoph Steinbeck. Review on natural products databases: where to find data in 2020. *Journal of Cheminformatics*, 12(1):20, December 2020.

[2] Ophélie Fraisier-Vannier, Justine Chervin, Guillaume Cabanac, Virginie Puech, Sylvie Fournier, Virginie Durand, Aurélien Amiel, Olivier André, Omar Abdelaziz Benamar, Bernard Dumas, Hiroshi Tsugawa, and Guillaume Marti. MS-CleanR: A Feature-Filtering Workflow for Untargeted LC–MS Based Metabolomics. *Analytical Chemistry*, 92(14):9971–9981, July 2020.

[3] François Chassagne, Guillaume Cabanac, Gilles Hubert, Bruno David, and Guillaume Marti. The landscape of natural product diversity and their pharmacological relevance from a focus on the Dictionary of Natural Products®. *Phytochemistry Reviews*, 18(3):601–622, June 2019.

Nouvelle signature pour le site GSEA à partir des métabolites

Syrine BOUALLEGUE[1] & Denis MESTIVIER[1]

[1] Plateforme de Bioinformatique , Institut Mondor de Recherche Biomédicale (IMRB/Inserm U955), Univ. Paris Est-Créteil, Faculté de Santé, 8 rue du Général Sarrail – 94010 Créteil, France

Corresponding Author: syrine.bouallegue@inserm.fr

L'analyse GSEA (Gene Set Enrichment Analysis) [1] est une étape incontournable en analyse transcriptomique (RNAseq) permettant d'organiser des sous-ensembles de gènes en fonctions biologiques, conduisant à une vision plus intégrative des jeux de données. La Molecular Signatures Database (MSigDB) offre une collection d'annotations d'ensemble de gènes (genesets) essentiellement basés sur données génomiques [2] ainsi qu'un environnement logiciel dédié.

En parallèle de ces informations, la banque de données « Human Metabolome Database » (HMDB) offre une source d'information complète sur les métabolites et le métabolisme chez l'homme [3]. Notre travail offre de lier ces deux sources d'information en effectuant une analyse GSEA/RNASeq.

Nous avons développé une signature Métabolomique en utilisant l'information d'association entre métabolites et gènes issus de la HMDB. A partir des 217 920 métabolites et 863 760 interactions Métabolite-Gène. Parmi les 19 178 genesets (comportant plus de 10 gènes) obtenus, plus de 97 % sont redondants et doivent être traités afin de ne pas biaiser les statistiques de correction pour des tests multiples par exemple. Après filtration, nous avons construit un geneset comportant 571 genesets « Métabolomiques ». Cette signature, au format "Gene Matrix Transposed" (GMT), permet de l'utiliser avec le logiciel GSEA du site (http://www.gsea-msigdb.org/gsea/index.jsp) et de bénéficier des tests statistiques et outils visuels associés. Note signature permet ainsi d'associer Transcriptomique et Métabolomique dans un premier niveau d'exploration multi-omics.

Nous présenterons l'illustration de notre signature Métabolomique aux données de RNAseq sur le jeu de données du Cholangiocarcinome du TCGA (TCGA-CHOL [4]) en comparant l'expression des gènes dans les tissus cancéreux versus les tissus normaux.

References

1. Subramanian, A. *et al.* Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. *Proc. Natl. Acad. Sci. U. S. A.* **102**, 15545–15550 (2005).
2. Kankainen, M., Gopalacharyulu, P., Holm, L. & Orešič, M. MPEA-metabolite pathway enrichment analysis. *Bioinformatics* **27**, 1878–1879 (2011).
3. Wishart, D. S. *et al.* HMDB 5.0: The Human Metabolome Database for 2022. *Nucleic Acids Res.* **50**, D622–D631 (2022).
4. Farshidfar, F. *et al.* Integrative Genomic Analysis of Cholangiocarcinoma Identifies Distinct IDH-Mutant Molecular Profiles. *Cell Rep.* **18**, 2780–2794 (2017).

DNA methylation profiling of ATM-deficient breast tumours

Nicolas Viart[1], Anne-Laure Renault[1,2], Sophia Murat El Houdigui[1], Séverine Eon-Marchais[1], Laetitia Fuhrmann[3], Dorothée Le Gal[1], Eve Cavaciuti[1], Marie-Gabrielle Dondon[1], Juana Beauvallet[1], Anne-Vincent Salomon[3], Dominique Stoppa-Lyonnet[4], Melissa Southey[2], Nadine Andrieu[1] and Fabienne Lesueur[1]

[1] Institut Curie, PSL University, INSERM U900, 75005, Paris, France
[2] Monash University, University of Melbourne, Cancer Council Victoria, Melbourne, Australia
[3] Service de Pathologie, Institut Curie, Paris, France
[4] Service de génétique, Institut Curie, INSERM U830, Université de Paris-Cité, Paris, France

Corresponding author: nicolas.viart@curie.fr

Germline bi-allelic inactivation of *ATM* causes ataxia-telangiectasia (A-T) disorder, characterized by genetic instability, radiosensitivity and predisposition to cancer. Women heterozygous for a rare predicted deleterious *ATM* variant have a two to four-fold increased risk of breast cancer (BC) as compared to non-variant carriers. However, the current imprecise estimates of BC risk associated with *ATM* variants prevents the establishment of management recommendations for women carrying an *ATM* variant and their relatives. Thus, no specific risk management strategies currently exist for these patients [1]. The identification of a molecular signature of ATM-deficient tumours would allow to better understand the aetiology of these cancers and to identify potential therapeutic targets. Recently, we showed that breast tumours developed by *ATM* deleterious variant carriers may display specific genomic alterations, notably bi-allelic inactivation of *ATM* and loss at the *RB1* locus [2]. Given that both ATM and its antagonist RB, regulate the function of the DNA methyltransferase DNMT1, we hypothesized that ATM-deficient tumours will have a distinctive genome-wide methylation pattern.

To address this question, we analysed DNA methylation profiles of 24 breast tumours from *ATM* variant carriers identified in A-T or in Hereditary Breast and Ovarian Cancer families. Nineteen patients carried a pathogenic or likely pathogenic variant and five patients carried a Variant of Unknown Significance (VUS). DNA methylation was measured using the Illumina Infinium HumanMethylationEPIC array and compared to a published dataset comprising 34 breast tumours of cases from the general population analysed with the same array. We first pre-processed raw data of the whole dataset using a custom pipeline developed with Minfi. We then mapped each probe of the array to genes and promoters using bedtools. We performed a gene set enrichment analysis (GSEA) with KEGG database and ClusterProfiler R package to search for pathways enriched in differentially methylated genes. We then used EnrichmentMap, a Cytoscape App, to group pathways according to gene similarities. Next, we performed a cluster analysis based on promoter methylation level for each group of genes.

We found 203 genes with differentially methylated promoters ($|\log_2(\text{fold change})|>1$ and adjusted p-value<0.05) when comparing tumours of carriers of a pathogenic or likely pathogenic *ATM* variant to tumours of cases from the general population. The GSEA detected 76 KEGG pathways significantly enriched in ATM-deficient tumours (p-value<0.05) with "Homologous recombination", "Fanconi anaemia pathway" and "p53 signalling pathway" among the 15 most enriched pathways. Furthermore, we identified three groups of overlapping pathways, containing 453, 298 and 77 genes, respectively. Methylation level of the group composed of 298 genes and regrouping cancer-related pathways, allowed to better discriminate ATM tumours: 13 out of the 19 tumours (68.4%) of pathogenic or likely pathogenic variant carriers were clustered. Hence, differentially methylated genes and enriched pathways may represent potential therapeutic targets. Additional GSEA using Gene Ontology and Reactome databases are underway to confirm these preliminary results. Next, tumour profiles of carriers of a VUS will be compared to these 19 ATM-deficient tumours to assess if methylation profile can be of use for variant classification. Transcriptomic analyses are also planned to assess the functional impact of methylation dysregulation occurring in ATM-deficient tumours and to complete the genomic profile of ATM tumours.

References

[1] Fabienne Lesueur, *et al.* First international workshop of the ATM and cancer risk group (4-5 December 2019). *Familial Cancer*, June 2021.

[2] Anne-Laure Renault, *et al.* Morphology and genomic hallmarks of breast tumours developed by ATM deleterious variant carriers. *Breast Cancer Research*, 20(1), 2018.

Deciphering the molecular network controlling the biology of the pig blastocyst and its cellular interactions

Adrien Dufour[1], Sarah Djebali[2] , Stephane Ferchaud[3] , Yoann Bailly[3] , Patrick Manceau[3] , Frederic Martins[4], Bertrand Pain[5], Sylvain Foissac[6], Jérôme Artus[7], Hervé Acloque[1*]

[1] Paris-Saclay University, INRAE, AgroParisTech, GABI, Jouy-en-Josas, France

[2] IRSD, Toulouse University, INSERM, INRAE, ENVT, Toulouse, France

[3] GenESI, UE 1372 Génétique, Expérimentations et Systèmes Innovants, 86480 Rouillé, France

[4] Toulouse Biotechnology Institute (TBI), Plateforme Genome et Transcriptome (GeT-Biopuces), Toulouse University, CNRS, INRAE, INSA, Toulouse, France

[5] University of Lyon 1, INSERM, INRAE, Stem Cell and Brain Research Institute, U1208, USC1361, Bron, France

[6] GenPhySE, Toulouse University, INRAE, ENVT, Castanet Tolosan, France

[7] INSERM U1310, Paris Saclay University, Villejuif, France

Corresponding Author : herve.acloque@inrae.fr

The embryonic development of the pig differs from that of humans and mice from the blastocyst stage and is characterised by much later implantation. This particular period is concomitant with a lengthening and a significant growth of the extraembryonic tissues of the embryo and is still poorly understood. These drastic changes occurring before implantation could thus affect the biology of embryonic pluripotent cells in a new way compared to our knowledge developed on primate or rodent cells.

To better understand the biology of pig embryos before implantation, we first produced a large dataset of single-cell RNAseq at different embryonic states (early, late, ovoid and elongated blastocysts). These data were cleaned, filtered and represent a total of 40,000 cells. With these data, we firstly characterised embryonic cellular population and their evolution, and we identified specific markers of these populations. We then inferred gene regulatory networks working on modules of gene regulation (regulon) using a pig adapted version of the SCENIC [1] package and selected those specifically active in each embryonic population. Meta-analysis on others scRNAseq publication on preimplementation embryo in pigs [2] and humans [3] enhance the confidence on our identified regulon. Our results confirm the molecular specificity of the three primary embryonic lineages (epiblast, trophectoderm and hypoblast) and identify stage-specific subpopulations. This allows us to infer the biological functions of these three main lineages and the interactions between them and identify key regulation modules linked to those functions. We also provide new insights into the biology of epiblast cells prior to implantation, and we observed a relatively constant pluripotent state in the epiblast over time for the studied stages. In addition, to discover chromatin specific landscape between the three main populations, a single-cell multi-omics dataset (paired scRNAseq and scATACseq) has been performed and will be briefly presented.

References

1. Aibar, et al. SCENIC: Single-Cell Regulatory Network Inference and Clustering. *Nature Methods* 14, n° 11, 2017.
2. Zhi, et al. Generation and Characterization of Stable Pig Pregastrulation Epiblast Stem Cell Lines. *Cell Research* 32, n° 4, 2022.
3. Meistermann, et al. Integrated Pseudotime Analysis of Human Pre-Implantation Embryo Single-Cell Transcriptomes Reveals the Dynamics of Lineage Specification. *Cell Stem Cell* 28, n° 9, 2021.

Modelling the dynamics of Salmonella infection in the gut at the bacterial and host levels

Coralie MULLER[1], Arie WORTSMAN[1], Pablo Andres UGALDE SALAS[1], Clémence FRIOUX[1], and Simon LABARTHE[2,1]

[1] Inria University of Bordeaux, 200 avenue de la Vieille Tour, 33405, Talence, France

[2] INRAe UMR BIOGECO, 69 route d'Arcachon, 33612, Cestas, France

Corresponding authors: coralie.muller@inria.fr, arie.wortsman@inria.fr

The human gut microbiota is a complex ecosystem composed of numerous bacteria that co-exist, forming communities that interact within each other but also with their host. These interactions occur essentially at the metabolic level, building a commensal relationship between the host and its microbiota through beneficial exchanges, metabolic niches and competition for nutrients. However, this balance can be disrupted by many factors, including infections by enteric bacteria such as *Salmonella*, a foodborne pathogen. To invade the anaerobic environment of the gut very favorable to commensals, enteric pathogens can trigger an inflammation in the epithelium, that initiates a cascade of shifts in the commensal microbiota and the host metabolism, ending up in the creation of a new aerobic metabolic niche in favor of the pathogen.

One way to understand the complex dynamics of these infections is to construct metabolic networks consisting of all the metabolic and molecular interactions that occur in an organism and use them into numerical models such as Flux Balance Analysis (FBA) [1]. FBA computes activity rates, or fluxes, in metabolic reactions of a bacteria alone or within a community. Dynamic FBA (dFBA) integrates the fluxes calculated in FBA into a system of ordinary differential equations (ODEs), enabling us to model the temporal dimension of metabolic simulations.

We model the host-microbiota-pathogen system by gathering three metabolic models : *Faecalibacterium prausnitzii* as a representative butyrate producing commensal bacterium, *Salmonella Enterica* serovar Thiphimurium for the enteric pathogen, these models are from the VMH database [2], and a metabolic model of human cell representative of the epithelial cell metabolism. Namely, we extracted and cleaned reactions occurring in colonocytes - epithelial cells of the colon - from a metabolic network of the whole human body [3], thereby creating a metabolic model representing the host part. The complexity of building this three-partner model came from the reconstruction of the colonocyte model and its connection to both the other networks as well as its environment through several compartments. The complete model consists in an ODE system that allows the exchange of metabolites between the host and the two bacteria, and the uptake and transfer of nutrients to blood compartments. To speed up the computation of the FBA models needed for extensive numerical exploration, the metabolic models are approximated by a metamodeling method [4]. The complete model correctly captures the dynamics of the complex cross-talk between the commensal, the pathogen and the host during infection.

The use of statistical learning in numerical metabolic modelling opens doors to more complex and previously time-consuming simulations, including spatio-temporal modelling.

References

[1] Ove Øyås and Jörg Stelling. Genome-scale metabolic networks in time and space. *Current Opinion in Systems Biology*, 8:51–58, 2018.

[2] Magnusdottir et al. Generation of genome-scale metabolic reconstructions for 773 members of the human gut microbiota. *Nature Biotechnology*, 35(1)::81–89, 2017.

[3] Ines Thiele, Swagatika Sahoo, Almut Heinken, Johannes Hertel, Laurent Heirendt, Maike K Aurich, and Ronan MT Fleming. Personalized whole-body models integrate metabolism, physiology, and the gut microbiome. *Molecular Systems Biology*, 16(5):1–24, 2020.

[4] Clémence Frioux, Sylvie Huet, Simon Labarthe, Julien Martinelli, Thibault Malou, David James Sherman, Marie-Luce Taupin, and Pablo Ugalde-Salas. Accelerating metabolic models evaluation with statistical metamodels: application to Salmonella infection models. working paper or preprint, April 2022.

Prioritization of Master Regulators Through Influence Maximization

Clémence RÉDA[1] and Andrée Delahaye-Duriez[1,2,3]
[1] Univ. Paris Cité, Neurodiderot, Inserm, F-75019 Paris
[2] Univ. Paris 13, Sorbonne Paris Nord, UFR de santé, médecine et biologie humaine, F-93000 Bobigny
[3] Unité fonct. de médecine génomique et génétique clinique, Hôp. Jean Verdier, AP-HP, F-93140 Bondy

Corresponding author: clemence.reda@inserm.fr

1 Introduction

Master regulator genes are at the top of the gene regulation hierarchy and allow to better understand regulatory dynamics. *In silico* detection of these genes, through system biology approaches, is a popular attempt at speeding up disease research. However, when considering a large number of genes, building a dynamic regulatory model becomes a tedious and time-consuming task. Moreover, current detection methods do not take into account regulatory cascades. Here, we describe a method to identify master regulatory genes, and apply it to find novel master regulator genes linked to epilepsy.

2 Methods

First, our work focus on combining public data sources to design an end-to-end pipeline for the synthesis of a dynamic gene regulatory network, starting from a subset of genes. This network models the regulatory dynamics in a well-chosen cell line. For epilepsy, we considered the gene module M30 associated with epileptic *de novo* mutations [1], and gene expression data from a neural progenitor and a neuroblastoma cell lines. Second, we defined the concept of "gene influence", in terms of transcriptomic impact of gene perturbation in this network. Finally, we applied an influence maximization algorithm [2] to retrieve genes with highest regulatory influence on the remainder of the network.

3 Results

Fig. 1 displays the Spearman's ρ correlation heatmap between influence values, network centrality measures (Control Centrality), and scores associated with the pathogenicity of genes (pLI, RVIS) in M30. Influence is consistent and strongly correlated with network-dependent measures. Moreover, we performed a over-representation analysis (ORA), which shows that top genes for influence (Fig. 2) are significantly enriched in epilepsy-related terms at level 5%, compared to the whole M30 module. This methodology allows a reproducible detection of master regulators, introducing for the first time a measure which takes into account transcriptional cascades.

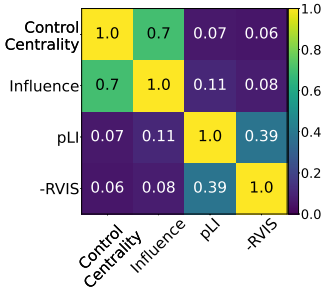


Fig. 1.

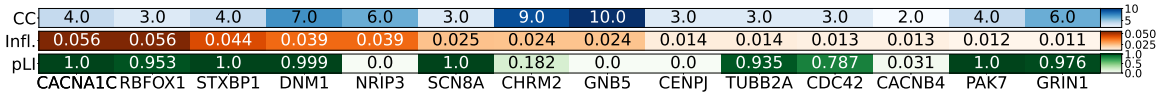


Fig. 2.

Acknowledgements

This work was supported by INSERM, the French Ministry of Higher Education and Research, Université Paris Cité, Université Sorbonne Paris Nord, and the French National Research Agency.

References

[1] Andree Delahaye-Duriez, Prashant Srivastava, Kirill Shkura, Sarah R Langley, Liisi Laaniste, Aida Moreno-Moral, Bénédicte Danis, Manuela Mazzuferi, Patrik Foerch, Elena V Gazina, et al. Rare and common epilepsies converge on a shared gene regulatory network providing opportunities for novel antiepileptic drug discovery. *Genome biology*, 17(1):1–18, 2016.

[2] David Kempe, Jon Kleinberg, and Éva Tardos. Maximizing the spread of influence through a social network. In *Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 137–146, 2003.

Using machine-learning on metabolomics data to predict complex phenotypes

Sylvain Prigent[1,2] and Yves Gibon[1,2]
[1] Univ. Bordeaux, INRAE, UMR1332 BFP, 33882 Villenave d'Ornon, France
[2] Bordeaux Metabolome, MetaboHUB, PHENOME-EMPHASIS, 33140 Villenave d'Ornon, France

Corresponding author: sylvain.prigent@inrae.fr

The metabolome is often seen as the ultimate cellular phenotype, resulting from all biochemical processes taking place in an organism. Using a robotized high throughout platform to prepare samples and modern mass spectrometers coupled with chromatography, it is nowadays possible to obtain both targeted and untargeted metabolomic data on hundreds of samples in a few days. Those experiments lead to thousands of metabolic features representing a precise description of the cell state at a given time. This state is a good proxy of the past, the present and the future of a given tissue, combining the result of the previous developmental stages, the current status and the bricks enabling the future development of the cells. Moreover, by applying different types of stresses to an organism, it is possible to increase the diversity of the metabolome, hence displaying broader characteristics.

By combining metabolomic and phenotypic data and using machine-learning approaches, we were able to accurately predict several complex phenotypes such as relative growth rate [1], grafting success [2], plant elevation [3] or resistance to pathogens [4] on various plant matrices. Besides, enabling prediction of those complex phenotypes, those techniques permit the discovery of biomarkers that give insights into mechanisms controlling the phenotypes.

Here we will present the different cases studied recently by our laboratory and metabolomic platform. The accuracies of the predictions and the physiological knowledge acquired thanks to those models will be detailed. Perspectives about the integration of knowledge acquired through the use of those machine-learning models into more mechanistic models will also be discussed.

Acknowledgements

The authors want to thank all partners involved in the acquisition of the data, the Genotoul bioinformatics platform Toulouse Occitanie (Bioinfo Genotoul, https://doi.org/10.15454/1.5572369328961167E12) for providing computing resources as well as the GenOuest bioinformatics core facility (https://www.genouest.org) for providing the computing infrastructure.

References

[1] Léa Roch, Sylvain Prigent, Holger Klose, Coffi-Belmys Cakpo, Bertrand Beauvoit, Catherine Deborde, Laetitia Fouillen, Pierre van Delft, Daniel Jacob, Björn Usadel, et al. Biomass composition explains fruit relative growth rate and discriminates climacteric from non-climacteric species. *Journal of Experimental Botany*, 71(19):5823–5836, 2020.

[2] Grégoire Loupit, Josep Valls Fonayet, Sylvain Prigent, Duyen Prodhomme, Anne-Sophie Spilmont, Ghislaine Hilbert, Céline Franc, Gilles De Revel, Tristan Richard, Nathalie Ollat, et al. Identifying early metabolite markers of successful graft union formation in grapevine. *Horticulture Research*, 2022.

[3] Thomas Dussarrat, Sylvain Prigent, Claudio Latorre, Stéphane Bernillon, Amélie Flandin, Francisca P Díaz, Cédric Cassan, Pierre Van Delft, Daniel Jacob, Kranthi Varala, et al. Predictive metabolomics of multiple atacama plant species unveils a core set of generic metabolites for extreme climate resilience. *New Phytologist*, 234(5):1614–1628, 2022.

[4] Estrella Luna, Amélie Flandin, Cédric Cassan, Sylvain Prigent, Chloé Chevanne, Camélia Feyrouse Kadiri, Yves Gibon, and Pierre Pétriacq. Metabolomics to exploit the primed immune system of tomato fruit. *Metabolites*, 10(3):96, 2020.

Multi-omic integration: adding network topology to study axial spondyloarthritis

Annabelle Beaudoin[1], Vincent Guillemot[1] and Natalia Pietrosemoli[1]
[1] Hub of Bioinformatics and Biostatistics, Institut Pasteur, 25-28 rue du Docteur Roux, 75015 Paris, France

Corresponding author: natalia.pietrosemoli@pasteur.fr

Abstract

We aim at improving current understanding of the pathophysiology of spondyloarthritis (SpA). Currently, there exist no treatments to cure this chronic inflammatory disease, and only limited therapies to treat the symptoms. Here, we analyze gene expression profiles and clinical for a cohort of 80 SpA patients whose cells have been exposed to two kinds of stimulation: Lipopolysaccharide stimulation (LPS), as a proxy of innate immunity and, Staphylococcal enterotoxin B stimulation (SEB), as a proxy of acquired immunity.

Here, we propose a method called netSGCCA, based on the integration of multiblock data. These types of approaches are now an essential tool to analyze the increasingly complex data that biologists, bioinformaticians and biostatisticians encounter on a daily basis: from multi-omics data, to imaging-genetic data. netSGCCA derives from framework of the Generalized Canonical Correlation Analysis, used to study the relationship between several groups of variables. In particular, we base our method on the SGCCA [1] (*Sparse Generalized Canonical Correlation Analysis*), which allows choosing the most pertinent variables when the blocks have a large number of variables, such as in omic-derived data. Our method uses the GraphNet penalty [2], offering the advantage of integrating network topology information reflecting the interactions among the variables within a given data block. netSGCCA may benefit from the rich and complex information present in biological reference databases such as STRING-DB, thus allowing to integrate known associations between the molecular players.

We apply netSGCCA to a study comprising three data blocks. Two blocks of gene expression data corresponding to the different stimulations (LPS and SEB) of 277 and 283 genes, respectively. One block of quantitative variables corresponding to three clinical progress scores. The reference network used was obtained from the Protein-Protein interactions STRING-DB database [3]. One of the objectives of the method is to remove the "high-frequency" components [4] induced by the GraphNet penalty, meaning sharp variations of the loadings between variables (e.g., genes) that are neighbors in the reference network. These high frequencies make difficult to interpret results. In order to solve this problem, we use a modified version of the graph's Laplacian.

netSGCCA offers promising results for integrating network topology information, further developments include how to define the optimal values for the different method parameters, including the GraphNet penalty.

References

[1] Arthur Tenenhaus, Cathy Philippe, Vincent Guillemot, Kim-Anh Le Cao, Jacques Grill, and Vincent Frouin. Variable selection for generalized canonical correlation analysis. *Biostatistics (Oxford, England)*, 15(3):569–83, jul 2014.

[2] Logan Grosenick, Brad Klingenberg, Kiefer Katovich, Brian Knutson, and Jonathan E. Taylor. Interpretable whole-brain prediction analysis with GraphNet. *NeuroImage*, 72:304–321, 2013.

[3] Damian Szklarczyk, Annika L. Gable, Katerina C. Nastou, David Lyon, Rebecca Kirsch, Sampo Pyysalo, Nadezhda T. Doncheva, Marc Legeay, Tao Fang, Peer Bork, Lars J. Jensen, and Christian von Mering. The STRING database in 2021: customizable protein-protein networks, and functional characterization of user-uploaded gene/measurement sets. *Nucleic acids research*, 49(D1):D605–D612, jan 2021.

[4] Franck Rapaport, Andrei Zinovyev, Marie Dutreix, Emmanuel Barillot, and Jean-Philippe Vert. Classification of microarray data using gene networks. *BMC bioinformatics*, 8:35, feb 2007.

Metamodelling of Dynamic Flux Balance Analysis

Clémence Frioux[1], Sylvie Huet[2], Simon Labarthe[1,3], Julien Martinelli[4,5], Thibault Malou[6], David Sherman[1], Marie-Luce Taupin[7] and Pablo Ugalde-Salas[1]

[1] Inria - Université de Bordeaux, 33400 Talence
[2] Université Paris-Saclay, INRAE, MaIAGE, 78350, Jouy-en-Josas, France
[3] INRAE, Univ. Bordeaux, BIOGECO, F-33610 Cestas
[4] INSERM U900, Saint-Cloud, France, Institut Curie, Saint Cloud, France, Paris Saclay University, France, MINES ParisTech, CBIO - Centre for Computational Biology, PSL Research University, Paris, France
[5] Lifeware Group, Inria Saclay Ile-de-France, Palaiseau 91120, France
[6] INSA - Institut de Mathématique de Toulouse, Toulouse, France
[7] Laboratoire LaMME, UEVE and UMR 8071, Université Paris Saclay, Evry, France

Corresponding author: pablo.ugalde-salas@inria.fr

1 Introduction

Flux balance analysis (FBA) allows us to predict the uptake and production rates of different metabolites and molecules of interest. FBA consists of a set of linear constraints describing the reactions encoded in the genome coupled with an objective function, thus creating a linear program (LP) very efficiently tackled by available solvers [1]. The nutritional environment which affects what and how much a cell can produce and grow is defined through the constraints of a FBA model. Dynamic flux balance analysis (dFBA) allows to consider the inherent temporal dimension of cells as entities immersed in a medium which changes through time either by the cell's own metabolic activity or by the flows of matter and energy through the environment. Community dFBA models can be obtained by coupling different dFBA models by their nutrient use.

Community dFBA consists of a set of differential equations with state variables that include both metabolites and cells populations concentrations through time, whose dynamics are described by reaction and exchange rates. Some of these rates are given by the state-dependent result of the FBA model of the different cell populations involved in the community. Numerically, solving a system of differential equations implies the computation of at least one FBA per population per time step. Even though LP are numerically very efficiently solved, a batch of LP problems can prove computationally expensive impairing computational exploration of the model [2].

In order to speed-up computations we propose a metamodelling technique based on Reproducing Kernel Hilbert Spaces (RKHS) that approximates the relationship between inputs and outputs of the FBA models. A learning database was assembled from the FBA outputs and nutrient concentrations obtained from several simulations of the dFBA. This input-output dependence is approximated by projection in an RKHS, which is done in an offline manner. We chose a specific RKHS space, the ANOVA-RKHS, which allows to perform variable selection, inducing additional speed-up [3].

We used a toy example based on the modelling of Salmonella infection of the colon where two FBA are involved in the differential equations. Replacing the FBA models by their metamodels in the community dFBA speeds up computations by a factor of 45, with a total relative error for the dFBA state variables maintained below 5%.

Acknowledgements

This study received fundings from Inria through the Exploratory Action SLIMMEST (for further in- formation see https://www.inria.fr/en/slimmest). Simon Labarthe got support for this study from the France-Berkeley Fund through the project Articulate.

References

[1] Jeffrey D Orth, Ines Thiele, and Bernhard Ø Palsson. What is flux balance analysis? *Nature Biotechnology*, 28(3):245–248, 2010.

[2] James D Brunner and Nicholas Chia. Minimizing the number of optimizations for efficient community dynamic flux balance analysis. *PLoS computational biology*, 16(9):e1007786, 2020.

[3] Sylvie Huet and Marie-Luce Taupin. Metamodel construction for sensitivity analysis. *ESAIM: Proceedings and Surveys*, 60:27–69, 2017.

Logic programs to infer computational models of the human embryonic development

Mathieu Bolteau[1], Jérémie Bourdon[1], Laurent David[2] and Carito Guziolowski[1]
[1] Nantes Université, École Centrale Nantes, CNRS, LS2N, UMR 6004, F-44000 Nantes, France
[2] Nantes Université, INSERM, Center for Research in Transplantation and Translational Immunology, UMR 1064, F-44000 Nantes, France

Corresponding author: mathieu.bolteau@univ-nantes.fr

1 Introduction

Define predictive models of complex biological systems is an important work in system biology. Answer-Set Programming (ASP) handles high combinatorial problems, e.g. to discriminate the response of Acute Myeloid Leukemia to treatment [1]. Here, our project is to develop a method to model a dynamic biological system using single-cell RNA-seq data: human pre-implantation development. *In vitro* fertilization (IVF) allows infertile couples to have babies, but works only 25% of IVF cycles are successful. Given limited access to human embryos, the field of IVF needs *in silico* models, in order to improve our understanding of embryo development. Sequential events drive the cell fate transitions during the human pre-implantation development. The first cell fate decision in the common fate morula segregates the EPI cells from early TE cells. Then, the early TE stage becomes late TE one during the TE maturation. Ultimately, our model would allow to predict how embryos respond to specific perturbation of the system, such as changes in the culture media composition.

2 Proposed method

Current machine learning approaches to handle single-cell data focus on limited gene sets, obtained through dimensionality reduction (e.g., Area Under Curve of cluster-specific genes) or gene-to-gene correlation (e.g., WGCNA). The data we use are single-cell RNA-seq data of cells extracted from different embryos at different development stages (representing the expression of $\sim 20,000$ genes for an atlas of $\sim 1,700$ cells from 128 embryos) [2]. The samples are precisely annotated: timelapse developmental stage and embryo of origin. Our objective is to infer Boolean networks (BNs) that will represent the human pre-implantation embryonic development. First, we will build a prior-knowledge network (PKN). We reconstruct our PKN, representing the gene interactions, using pyBRAvo [3] which automatically assembles gene regulatory networks using Web Semantic tools. Second, we will use the caspo software [4], which learn BNs from the PKN combined to experimental responses of perturbations inferred from the single-cell RNA-seq data. The main challenge is to emulate a perturbation from a time series, without perturbation. We aim to use the different fates (e.g., EPI vs TE) or stages (e.g., early TE vs late TE) as pseudo-perturbation to infer BNs using ASP. Altogether, our model will greatly contribute to improve IVF cycles.

Acknowledgements

This work was co-supported by ANR AIBY4 (ANR-20-THIA-0011) and ANR BOOSTIVF (ANR-20-CE17-0007) projects.

References

[1] Lokmane Chebouba et al. Discriminate the response of acute myeloid leukemia patients to treatment by using proteomics data and answer set programming. *BMC Bioinformatics*, 19:15–26, 3 2018.

[2] Dimitri Meistermann et al. Integrated pseudotime analysis of human pre-implantation embryo single-cell transcriptomes reveals the dynamics of lineage specification. *Cell Stem Cell*, 28:1625–1640.e6, 9 2021.

[3] M. Lefebvre et al. Large-scale regulatory and signaling network assembly through linked open data. *Database*, 2021, 10 2021.

[4] Santiago Videla et al. caspo: a toolbox for automated reasoning on the response of logical signaling networks families. *Bioinformatics*, 33:947–950, 3 2017.

Machine Learning classification performance on mechanistic representations of the gut microbiota built from abundance profiles

Baptiste Ruiz[1], Arnaud Belcour[1], Samuel Blanquart[1], Isabelle Le Huërou-Luron[2], Sylvie Buffet-Bataillon[3], Yann Le Cunff[1] and Anne Siegel[1]

[1] Univ Rennes, Inria, CNRS, IRISA, Campus de Beaulieu, 263 Av. Général Leclerc, 35042 Rennes, France

[2] Institut NuMeCan, INRAE, INSERM, Univ Rennes, Saint-Gilles, France

[3] Department of Clinical Microbiology, CHU Rennes, Rennes, France

Corresponding Author: baptiste.ruiz@inria.fr

1 Abstract

Understanding the gut microbiota and its mechanisms has become a major point of interest in the medical field, with more and more studies correlating it to a variety of pathologies [1]. Machine Learning methods have been applied to this issue, approaching the microbiome as a predictor of the subjects' health [2-5]. These approaches however have yet to tap into the potential augmentation of the microbiome data which could be achieved by gathering information correlated to the microbiota's composition. In particular, the recognised micro-organisms' functional annotations have been suggested as a promising lead to enhance the comprehension of the microbiota as a metabolic network [6,7]. In line with this approach, we propose a new method to shift the representation of the gut microbiota from relative OTU abundances to a numeric mapping of the associated functional annotations, creating a mechanistic description of the microbial community. We have then explored the performances of Random Forest classifiers, a classic Machine Learning approach for microbiota classification, when applied to data converted to this new paradigm. This led to us finding that for a small sacrifice in classification performance, this approach could help highlight important metabolic mechanisms. Exploiting this method would also yield more thorough and complete results than what can be gathered through the standard approach based on finding OTUs that make a difference between classes of subjects.

References

[1] Cho, I., Blaser, M. The human microbiome: at the interface of health and disease. *Nat Rev Genet* **13,** 260–270 (2012). https://doi.org/10.1038/nrg3182

[2] Dan Knights, Elizabeth K. Costello, Rob Knight, Supervised classification of human microbiota, *FEMS Microbiology Reviews*, Volume 35, Issue 2, March 2011, Pages 343–359, https://doi.org/10.1111/j.1574-6976.2010.00251.x

[3] Statnikov, A., Henaff, M., Narendra, V. *et al.* A comprehensive evaluation of multicategory classification methods for microbiomic data. *Microbiome* **1,** 11 (2013). https://doi.org/10.1186/2049-2618-1-11

[4] Pasolli E, Truong DT, Malik F, Waldron L, Segata N (2016) Machine Learning Meta-analysis of Large Metagenomic Datasets: Tools and Biological Insights. PLOS Computational Biology 12(7): e1004977. https://doi.org/10.1371/journal.pcbi.1004977

[5] Oh, Min, and Liqing Zhang. "DeepMicro: deep representation learning for disease prediction based on microbiome data." Scientific reports 10.1 (2020): 1-9.

[6] Douglas, G.M., Hansen, R., Jones, C.M.A. *et al.* Multi-omics differentially classify disease state and treatment outcome in pediatric Crohn's disease. *Microbiome* **6,** 13 (2018). https://doi.org/10.1186/s40168-018-0398-3

[7] Casey M A Jones, MSc, Jessica Connors, PhD, Katherine A Dunn, PhD, Joseph P Bielawski, PhD, André M Comeau, PhD, Morgan G I Langille, PhD, Johan Van Limbergen, MD, FRCPCH, PhD, Bacterial Taxa and Functions Are Predictive of Sustained Remission Following Exclusive Enteral Nutrition in Pediatric Crohn's Disease, *Inflammatory Bowel Diseases*, Volume 26, Issue 7, July 2020, Pages 1026–1037, https://doi.org/10.1093/ibd/izaa001

What are the functions of the short open reading frame-encoded peptides in monocytes? An interactomic approach.

Sébastien A. CHOTEAU[1,2], Philippe PIERRE[2], Lionel SPINELLI[1], Andreas ZANZONI[1], Christine BRUN[1,3]

[1] Aix-Marseille Univ, INSERM, TAGC, CENTURI, Marseille, France

[2] Aix-Marseille Univ, CNRS, INSERM, CIML, CENTURI, Marseille, France

[3] CNRS, Marseille, France

Corresponding Authors: sebastien.choteau@univ-amu.fr, christine-g.brun@inserm.fr

Over the past decades, hundreds of thousands of short open reading frames (sORFs) have been identified on most eukaryotic RNAs, including long non-coding RNAs. Some of these ubiquitous elements are conserved across species and may eventually encode functional peptides [1]. Most of these peptides, called sORF-encoded peptides (sPEPs), have failed to be annotated notably due to the short length of their ORF (< 100 codons) and the use of alternative start codons (other than AUG). So far, the roles of only few sPEPs have been characterized. sPEPs whose function has been determined seem to be involved in a wide range of key biological processes such as apoptosis, mitochondrial complexes integrity, DNA reparation, mTOR signaling, transcriptional regulation, antigen presentation in eukaryotes or cardiac activity regulation in *D. melanogaster* [2]. This broad range of functions and their roles in physiological functions suggests sPEPs may constitute a new pool of therapeutic targets. Nonetheless, most of sPEPs remain unknown or poorly characterized.

In order to scrutinize their roles, we are studying the way sPEPs interact with canonical proteins, whose functions are known. To that extent, we gathered and homogenized publicly available data characterizing the human sORFs into a database, MetamORF (https://metamorf.hb.univ-amu.fr) [3], that has been exploited to study the functions of sPEPs. 10,475 sPEPs encoded by sORFs identified in monocytes have been recovered from this database. The interactions between these sPEPs and the canonical proteins expressed in monocytes have then been inferred using mimicINT, a method we developed. mimicINT is a computational method that allows to predict protein-protein interactions. It identifies the short linear motifs (SLiMs) and the globular domains on sPEPs and canonical proteins, two major protein interaction interfaces. SLiMs are short stretches of 3-10 contiguous amino acids, usually located in disordered regions [4]. mimicINT then uses experimentally validated patterns of SLiM-domain and domain-domain interactions to infer a network of sPEP-protein interactions. Finally, Monte-Carlo simulations are used to assess the SLiM functionality by estimating the likelihood of each SLiM observed in the sequences to occur by chance, which allows to discard the most likely false positive interactions from the network inferred.

This first sPEP-canonical proteins human interactome in monocytes contains nearly 950,000 binary interactions. We are currently exploring it to understand in which biological processes and signaling pathways sPEPs are involved. Interestingly, our preliminary analyses showed that, in monocytes, the sPEPs that are encoded on genes annotated as involved within a biological process are preferentially targeting the genes of this same biological process (BP) for most of the BPs (significant enrichment for 68% BP generic GO terms).

Acknowledgements

S.A.C. is funded by a Fondation pour la Recherche Médicale “Espoirs de la recherche” fellowship (FDT202106013072). Centre de Calcul Intensif d’Aix-Marseille is acknowledged for granting access to its high performance computing resources.

References

1. Pueyo *et al*. New Peptides Under the s(ORF)ace of the Genome. *Trends Biochem Sci*. 41(8):665-678, 2016.
2. Plaza *et al.* In search of lost small peptides. *Annu. Rev. Cell Dev. Biol.*, 33, 391–416, 2017.
3. Choteau *et al*. MetamORF: a repository of unique short open reading frames identified by both experimental and computational approaches for gene and metagene analyses. *Database*, baab032, 2021.
4. Zanzoni *et al.* Perturbed human sub-networks by Fusobacterium nucleatum candidate virulence proteins. Microbiome 5, 89, 2017.

A new *Penicillium chrysogenum* Genome Scale-Metabolic Network: reconciliation of previous data and focus on specialised metabolism

Delphine NEGRE[1,2], Abdelhalim LARHLIMI[2] and Samuel BERTRAND[1]

[1] Nantes Université, Institut des Substances et Organismes de la Mer, ISOMer, UR 2160, F-44000 Nantes, France.

[2] Nantes Université, École Centrale Nantes, CNRS, LS2N, UMR 6004, F-44000 Nantes, France

Corresponding Author: delphine.negre@univ-nantes.fr

Filamentous fungi are known to be reservoirs of specialised metabolites and much knowledge about their metabolism has been accumulated over time. Genome exploration of these organisms has revealed the high proportion of specialised metabolites produced by enzymes related to *biosynthetic gene clusters*. Modelling metabolism through reconstruction of Genome-Scale Metabolic Network (GSMN) is one way to understand their biosynthesis mechanisms.

Since the genome sequencing of *Penicillium chrysogenum* Wisconsin 54-1255 in 2008, several GSMNs reconstructions have been made [1, 2]. However, the rapid evolution and multiplication of data make the comparison or reuse of these models tricky.

Following current convention standards and quality criteria, the new reconstruction that we propose results from the following four elements. In parallel, *(1)* an update functional annotation of the *P. chrysogenum* genome was carried out and then supplemented by *(2)* an orthology search with different GSMNs models. This first draft was then enriched *(3)* by integrating data from previous GSMN reconstructions of *P. chrysogenum* and complemented *(4)* by manual curation steps targeting basal and specialised metabolism (supported by bibliographic data and by metabolomic data obtained in the laboratory). Of the 233 metabolites monitored, 204 are topologically producible. Keeping traceability of these different steps allows us to apprehend the complementarity of the approaches used as well as their relevance. Identifiers enrichment, based on MetaNetX [3], ensures interoperability between various databases [4].

Thus, the proposed *high-quality GSMN* has a MEMOTE [5] *score* of 67% and a *metabolic coverage* of 43%. In comparison, the networks published in 2013 and 2018 have a score of 31% and 16% respectively and offer a metabolic coverage of 8% and 14%. The model is composed of 5,191 metabolites interconnected by 5,888 reactions of which 5,025 are at least supported by a genomic sequence. *In fine*, GSMN exploration, with various physical and biological constraints, is expected to allow us to get insights into the natural products production and their precursors.

Keywords: Genome-Scale Metabolic Network (GSMN), data integration, interoperability, reconciliation

Acknowledgements

This work was supported by the ANR-18-CE43-0013 FREE-NPs.

References

1. Agren R, Otero JM, Nielsen J. Genome-scale modeling enables metabolic engineering of *Saccharomyces cerevisiae* for succinic acid production. J Ind Microbiol Biotechnol. 2013;40:735–47.
2. Prigent S, Nielsen JC, Frisvad JC, Nielsen J. Reconstruction of 24 *Penicillium* genome-scale metabolic models shows diversity based on their secondary metabolism. J Biochem Microbiol Technol. 2018;115:2604–12.
3. Moretti S, Martin O, Van Du Tran T, Bridge A, Morgat A, Pagni M. MetaNetX/MNXref – reconciliation of metabolites and biochemical reactions to bring together genome-scale metabolic networks. Nucleic Acids Res. 2016;44:D523–6.
4. Pham N, Van Heck RGA, van Dam JCJ, Schaap PJ, Saccenti E, Suarez-Diez M. Consistency, Inconsistency, and Ambiguity of Metabolite Names in Biochemical Databases Used for Genome-Scale Metabolic Modelling. Metabolites. 2019;9:28.
5. Lieven C, Beber ME, Olivier BG, Bergmann FT, Ataman M, Babaei P, *et al.* MEMOTE for standardized genome-scale metabolic model testing. Nat Biotechnol. 2020;38:272–6.

Met4J, a programmatic toolbox for graph-based analysis of metabolic networks

Ludovic Cottret[1] and Clément Frainay[2]
[1] LIPME (Laboratoire des interactions plantes - microbes - environnement), INRAE, CNRS, Université de Toulouse, Castanet-Tolosan, France
[2] Toxalim (Research Center in Food Toxicology), INRAE, Université de Toulouse, ENVT, INP-Purpan, UPS, Toulouse, France

Corresponding author: ludovic.cottret@inrae.fr

1 Context

Graph-theory based methods are essential tools for network analysis in various domains, including biology. Despite successful applications to metabolic networks over the last decades, including several developments specific to these models, few off-the-shelf implementations are openly available and easily findable. Furthermore, the conversion from metabolic model to meaningful graph is far from straightforward, yet tools for such endeavour are scarce. Beyond the necessary tailored pre-processing of models, the SBML interchange format[1] adopted for the most accurate models is incompatible with main generic graph-analysis libraries.

2 Description

We present met4J, an open-source library dedicated to the structural analysis of metabolism and graph-based analysis of metabolism-related data. The library allows to create multiple graph types encompassing metabolic models content, as well as several utilities for model scrapping and edition.

Met4J also includes a toolbox gathering implementations with command line interface of comprehensive pipelines for several analyses relevant to metabolism-related research, including network expansion[2], MetaboRank[3] and metabolic route search[4]. Some analyses will also be made available in a web browser through the MetExplore[5] webserver and a Galaxy instance.

3 Availability

Met4J source code and toolbox, proposed as executable JAR and docker or singularity images, are available at https://forgemia.inra.fr/metexplore/met4j under CeCILL open license. Library artifact is accessible through maven central repository for inclusion in other projects.

References

[1] M Hucka, A Finney, H M Sauro, H Bolouri, J C Doyle, H Kitano, B J Bornstein, D Bray, A A Cuellar, S Dronov, E D Gilles, M Ginkel, V Gor, I I Goryanin, W J Hedley, T C Hodgman, P J Hunter, N S Juty, J L Kasberger, A Kremling, U Kummer, L M Loew, D Lucio, P Mendes, E Minch, E D Mjolsness, Y Nakayama, M R Nelson, P F Nielsen, T Sakurada, J C Schaff, B E Shapiro, T S Shimizu, H D Spence, J Stelling, K Takahashi, M Tomita, J Wagner, and J Wang. The systems biology markup language (SBML): a medium for representation and exchange of biochemical network models. 19(4):524–531, 2003.

[2] Thomas Handorf, Oliver Ebenhöh, and Reinhart Heinrich. Expanding Metabolic Networks: Scopes of Compounds, Robustness, and Evolution ¨. *Journal of molecular evolution*, 61(4):498–512, 2005.

[3] Clément Frainay, Sandrine Aros, Maxime Chazalviel, Thomas Garcia, Florence Vinson, Nicolas Weiss, Benoit Colsch, Frédéric Sedel, Dominique Thabut, Christophe Junot, and Fabien Jourdan. MetaboRank : network-based recommendation system to interpret and enrich metabolomics results. (July), 2018.

[4] Didier Croes, Fabian Couche, Shoshana J Wodak, and Jacques Van Helden. Inferring meaningful pathways in weighted metabolic networks. *Journal of Molecular Biology*, 356(1):222–236, 2006.

[5] Ludovic Cottret, Clément Frainay, Maxime Chazalviel, Floréal Cabanettes, Yoann Gloaguen, Etienne Camenen, Benjamin Merlet, Stéphanie Heux, Jean-Charles Portais, Nathalie Poupin, and Others. MetExplore: collaborative edition and exploration of metabolic networks. *Nucleic acids research*, 46(W1):W495—-W502, 2018.

Improving the analysis of toxicants Mechanisms of Action with condition-specific models and network analysis

Louison Fresnais[1,2], Olivier Perin[2], Anne Riu[2], Clément Frainay[1], Fabien Jourdan[1] and Nathalie Poupin[1]

[1] Toxalim (Research Centre in Food Toxicology), Université de Toulouse, INRAE, ENVT, INP-Purpan, UPS, Toulouse, France.

[2] Digital Sciences Department, L'Oréal Advanced Research, Aulnay-sous-bois, France.

Corresponding author: louison.fresnais@rd.loreal.com

1 Abstract

The 3R rule strongly encourages reducing the use of *in vivo* studies, which are even forbidden in some research fields such as cosmetic. As a result, it is crucial to develop new methods to understand the Mechanisms of Action(MoA) of potentially hepatotoxic molecules. Current read-across strategies, which were first developed as potential *in silico* alternatives large-scale evaluation methods, are known to have some flaws (e.g indirect representation of metabolism, genes considered independents and post-translational modifications ignored). To overcome these limitations, we developed a new system biology approach based on Genome-Scale Metabolic Networks(GSMN), which, thanks to its topology, allows for a direct representation of metabolism and post-translational modifications, as well as dependencies between genes, proteins, reactions. Combining this new source of information with existing ones (e.g., transcriptomics, structural data, metabolomics, etc) should improve our understanding of hepatotoxic molecules MoA, and by extension, our ability to predict the toxicity of new molecules. To achieve these objectives, we developed a workflow that can be divided into two parts: condition-specific modelling and knowledge extraction by network analysis. The aim of the modelling part is to combine transcriptomic data and metabolic network topology to generate metabolic networks that are representative of the metabolic impact of a chemical compound. It starts with the pre-processing and binarization of transcriptomic data, which is then integrated by constraint-based modelling into the metabolic network using iMAT[1], to find the activated part of the network in adequation with transcriptomic data. However, usually one of the many possible solutions is arbitrarily picked.To tackle this, we implemented an extra enumeration step, inspired from DEXOM[2], which finds thousands of different but equally optimal condition-specific models, improving the robustness of the results. The network analysis step then extracts meaningful information from all the condition-specific models to understand the MoA of the studied molecule. In this step, we first extract reactions which are differentially activated after exposure to the molecule, and then we compute the pairwise distance matrix of the differentially activated reactions, based on the shortest path in the GSMN. This distance matrix is used to identify specific clusters of reactions in the metabolic network that can be extracted into small subnetworks. To facilitate the biological interpretation, these subnetworks can be analyzed and visualized using MetExploreViz[3]. Finally, combining thousands of condition-specific models representing a molecule's effect in specific metabolic subnetworks should improve our understanding of toxicants MoA. As a use case, we applied this workflow to model the effect of Amiodarone on PHH using transcriptomics data from Open TG-GATES[4]

References

[1] Tomer Shlomi, Moran N Cabili, Markus J Herrgård, Bernhard Ø Palsson, and Eytan Ruppin. Network-based prediction of human tissue-specific metabolism. *Nat. Biotechnol.*, 26(9):1003–1010, sep 2008.

[2] Pablo Rodríguez-Mier, Nathalie Poupin, Carlo de Blasio, Laurent Le Cam, and Fabien Jourdan. DEXOM: Diversity-based enumeration of optimal context-specific metabolic networks. *PLOS Comput. Biol.*, 17(2):e1008730, feb 2021.

[3] Maxime Chazalviel, Clément Frainay, Nathalie Poupin, Florence Vinson, Benjamin Merlet, Yoann Gloaguen, Ludovic Cottret, and Fabien Jourdan. MetExploreViz: web component for interactive metabolic network visualization. *Bioinformatics*, 34(2):312–313, jan 2018.

[4] Yoshinobu Igarashi, Noriyuki Nakatsu, Tomoya Yamashita, Atsushi Ono, Yasuo Ohno, Tetsuro Urushidani, and Hiroshi Yamada. Open TG-GATEs: a large-scale toxicogenomics database. *Nucleic Acids Res.*, 43(Database issue):D921–D927, jan 2015.

The UNTWIST project: Network analysis and performance modelling of *Camelina sativa* under thermal and water stress

Malo Le Boulch[1], Cédric Cassan[1,2], Pierre Pétriacq[1,2], Yves Gibon[1,2] and Sylvain Prigent[1,2]
[1] Univ. Bordeaux, INRAE, UMR1332 BFP, 33882 Villenave d'Ornon, France
[2] Bordeaux Metabolome, MetaboHUB, PHENOME-EMPHASIS, 33140 Villenave d'Ornon, France

Corresponding author: malo.le-boulch@inrae.fr

Climate change, especially climatic variability, poses critical challenges to European agriculture causing drought and high temperature stress to crops, resulting in productivity and yield loss. *Camelina sativa* is a native traditional European oilseed crop which has regained some attention thanks to its adaptability, yield stability and high performance in variable environments [1]. Camelina lipids in particular are a subject of research as they have multiple uses (feed, biofuel and green chemistry) [2]. The UNTWIST project will unravel the stress response mechanisms of Camelina in a multidisciplinary scientific project.

UNTWIST generates an unprecedented dataset, coming from a core collection of 54 *Camelina sativa* lines growing under different stresses and locations across Europe in greenhouses and fields. This data will be processed in order to link performance of Camelina to phenotypic traits, by classical univariate and multivariate statistical analysis, and in a second phase through a double modeling approach (top-down and bottom-up modelling).

The top-down modelling aims to predict phenotypic field traits from metabolic data. This involved creating a predictive model based on machine learning by combining phenotypic field data with metabolic data coming from targeted and untargeted metabolomics of early stage leaves of core collection Camelina lines grown in greenhouse under control condition and water or thermal stress. This data will be supplemented by redox and carbon isotope data from the same plants to achieve the predictive model. This model will be assessed against plants of Camelina coming from future field experiments. Once validated, metabolic variables with a strong positive or negative effect on phenotypic traits will be selected and annotated to find the underlying metabolites. The first results are encouraging, especially the correlation between metabolic profile and thousand kernel weight, with a correlation between 0.53 and 0.77 depending on the stress and the location of the field.

The bottom-up approach will focus on the growth and the development of the fruit (seed and silique) by the reconstruction of four compartmentalized genome scale metabolic networks of four focus lines of Camelina that have been identified as representing the diversity of answers to stress, based on their genomic data. These networks will be later refined with transcriptomics, proteomics and DNA methylation data from the four focus lines. The metabolic data used for the top-down modelling will be used to calculate input and output fluxes of each model in order to constrain networks to give insights into the mechanism involved in drought and heat tolerance.

References

[1] Marisol Berti, Russ Gesch, Christina Eynck, James Anderson, and Steven Cermak. Camelina uses, genetics, genomics, production, and management. 94:690–710.

[2] Jean-Denis Faure and Mark Tepfer. Camelina, a swiss knife for plant lipid biotechnology. 23(5):D503. Number: 5 Publisher: EDP Sciences.

Logical Modeling of Dysferlinopathies

Nadine BEN BOINA[1,2], Brigitte MOSSÉ[1], Anaïs BAUDOT[2] and Elisabeth REMY[1]

[1] Aix Marseille Univ, CNRS, I2M, Marseille, France

[2] Aix Marseille Univ, INSERM, MMG, CNRS, Marseille, France

Corresponding Author: Nadine.ben-boina@univ-amu.fr

Context

Dysferlinopathies are a set of rare muscle diseases caused by mutations in the dysferlin gene (*DYSF*) [1]. The symptoms are severe weakness and atrophy of skeletal muscles. Physiopathology shows defects in muscle cell membrane repair, abnormal muscle fiber regenerations alongside increased inflammatory response, and the presence of necrotic fibers [1]. Dysferlin is directly involved in the repair of cells membrane [2]. Notably, mutations in *DYSF* can cause those symptoms either in proximal or distal muscles [1]. To the best of our knowledge, there is no explanation as to how similar mutations in *DYSF* can lead to variation in symptoms.

Purpose

The aim of this work is to better understand the complex genotype-phenotype relationships of dysferlinopathies through the study of the regulatory networks of its disease gene. More precisely, we propose to build a logical model of *DYSF*-associated cellular processes, focusing on the functions impacted in dysferlinopathies. We aim to understand the mechanisms of control of the dynamical behaviors (at the cellular components level) and the appearance of phenotypic heterogeneities specific to these diseases.

Methods

Briefly, logical models consist in regulatory graphs in which nodes can represent any cellular component, including genes, proteins, or even a phenotypic state. Each node is associated with a discrete variable corresponding to a binary abstraction of its level of activity. Directed edges between nodes represent regulatory interaction (i.e., inhibitions or activations). The evolution of the activity level of each node is defined by a logical regulatory function. Biological information required to construct the model are obtained from literature and omics data. Overall, the state of the model corresponds to the activity level of all the nodes. The transition from one state to the next is given by the set of logical functions, and the resulting trajectories are stored in a state transition graph (STG). As the system evolves, it reaches attractors of the STG, which are either stable states (attractor of size one), or complex attractors (set of states in which the system cycles indefinitely).

Results and Perspectives

We built a *phenomenological* model in order to have a broad understanding of the main biological processes and cellular components involved in the diseases, and of their cross-talks. Through the inhibition of the membrane repair function, we were able to model dysferlinopathies physiopathology. Indeed, the attractors of the model (with defective membrane repair) can be assimilated with the phenotype of dysferlinopathies. We are now completing the model at the molecular level to obtain a Boolean model reflecting the phenotypic heterogeneities associated with *DYSF* mutations. In parallel, we are developing a theoretical analysis of the impact of the updating rules (used to construct the STG) on the dynamics of the Boolean models. We are analyzing in-depth three updating approaches (synchronous, asynchronous, most permissive). Our goal is to select the most suitable updating rule for the modelling of dysferlinopathies physiopathology.

References

1. Fanin M, Angelini C. Muscle pathology in dysferlin deficiency. Neuropathol Appl Neurobiol. 2002 Dec;28(6):461-70. doi: 10.1046/j.1365-2990.2002.00417.x. PMID: 12445162.
2. Bansal D, Miyake K, Vogel SS, Groh S, Chen CC, Williamson R, McNeil PL, Campbell KP. Defective membrane repair in dysferlin-deficient muscular dystrophy. Nature. 2003 May 8;423(6936):168-72. doi: 10.1038/nature01573. PMID: 12736685.

Towards a data-driven network inference of interactions between immune and cancer cells in Chronic Lymphocytic Leukemia

Hugo Chenel[1,2], Malvina Marku[1,2], Julie Bordenave[1,2], Nina Verstraete[1,2], Leila Khajavi[3] and Vera Pancaldi[1,2,4]

[1] INSERM, Cancer Research Center of Toulouse, 2 Avenue Hubert Curien, 31037, CEDEX 1 Toulouse, France.
[2] Université Toulouse-III Paul Sabatier, Route de Narbonne, 31330 Toulouse, France.
[3] Institut Universitaire du Cancer de Toulouse-Oncopole, 31330 Toulouse, France.
[4] Barcelona Supercomputing Center, Carrer de Jordi Girona, 29, 31, 08034 Barcelona, Spain.

Corresponding author: hugo.chenel@inserm.fr, malvina.marku@inserm.fr

Immune cells play a major role in cancer development. The characterization of these cells allows a better understanding of the tumor microenvironment (TME), a complex system containing multiple cell types interacting through contact and cytokine exchanges. Macrophages can be broadly categorized into two polarization states: M1, pro-inflammatory, or M2, anti-inflammatory and involved in wound healing. Interestingly, macrophages infiltrating tumors can be educated by cancer cells to promote tumor growth, thus becoming M2-like tumor-associated macrophages (TAMs). TAMs secrete cytokines and chemokines, exerting an anti-apoptotic, proliferative and pro-metastatic effect on tumor cells. This type of macrophages is present in many cancers, including chronic lymphocytic leukemia (CLL), a disease characterized by increased production of mature but dysfunctional B lymphocytes. Current CLL treatments target the formation of these TAMs in patient lymph nodes, remodeling the TME to arrest CLL cell proliferation [1]. Transcriptomics time courses on CLL patient blood allow studying the gene regulatory networks and interactions between immune and cancer cells in vitro to characterize the formation of TAMs. Inferring regulatory networks from transcriptomics data is an evolving subject, with several methods and inference algorithms that use bulk [2] or single-cell [3] gene expression data, based on steady state data or time series to uncover gene interactions [4]. Inferring such networks from data allows us to discover new molecular interactions, potentially identifying new drug targets.

The goal of this project is to use gene regulatory networks to investigate the crosstalk between macrophages and CLL cells. We perform gene regulatory network inference on both CLL and macrophage cells, using the dynGENIE3 [5] inference method on a unique transcriptomics time-series on purified macrophages and CLL cells from a 13-days co-culture. We focus on Transcription Factors (TF) only and applied network analysis tools to explore how the structural features of the network give information on the importance of the TFs included in the macrophage and CLL networks. Enrichment analysis of the TFs in the inferred networks shed light on the processes taking place inside the two cell populations. We are currently exploring how to integrate these entirely data-driven results with information from databases and literature based models to turn these networks into executable logical models.

References

[1] Oana Mesaros, Laura Jimbu, Alexandra Neaga, Cristian Popescu, Iulia Berceanu, Ciprian Tomuleasa, Bogdan Fetica, and Mihnea Zdrenghea. Macrophage polarization in chronic lymphocytic leukemia: Nurse-like cells are the caretakers of leukemic cells. *Biomedicines*, 8(11):516, 2020.

[2] Daniel Marbach, James C. Costello, Robert Küffner, Nicci Vega, Robert J. Prill, Diogo M. Camacho, Kyle R. Allison, Manolis Kellis, James J. Collins, and Gustavo Stolovitzky. Wisdom of crowds for robust gene network inference. *Nature methods*, 9(8):796–804, 2012.

[3] Christopher A Jackson, Dayanne M Castro, Giuseppe-Antonio Saldi, Richard Bonneau, and David Gresham. Gene regulatory network reconstruction using single-cell RNA sequencing of barcoded genotypes in diverse environments. *eLife*, 9:e51254, 2020.

[4] Chao Sima, Jianping Hua, and Sungwon Jung. Inference of gene regulatory networks using time-series data: A survey. *Current Genomics*, 10(6):416–429, 2009.

[5] Vân Anh Huynh-Thu and Pierre Geurts. dynGENIE3: dynamical GENIE3 for the inference of gene networks from time series expression data. *Scientific Reports*, 8(1):3384, 2018.

An agent-based model of tumor-associated macrophage differentiation in chronic lymphocytic leukemia

Nina VERSTRAETE[1, 2], Malvina MARKU[1, 2], Marcin DOMAGALA[1, 2], Hélène ARDUIN[1, 2], Julie BORDENAVE[1, 2], Jean-Jacques FOURNIÉ[1, 2], Loïc YSEBAERT[1, 2, 4], Mary POUPOT[1, 2] and Vera PANCALDI[1, 2, 3]

[1] INSERM, Cancer Research Center of Toulouse, 2 Avenue Hubert Curien, 31037, CEDEX 1 Toulouse, France

[2] Université Toulouse-III Paul Sabatier, Route de Narbonne, 31330 Toulouse, France

[3] Barcelona Supercomputing Center, Carrer de Jordi Girona, 29, 31, 08034 Barcelona, Spain

[4] Service d'Hématologie, Institut Universitaire du Cancer de Toulouse-Oncopôle, 31330 Toulouse, France

Corresponding Authors: nina.verstraete@inserm.fr, vera.pancaldi@inserm.fr

In the tumor microenvironment, tumor-associated macrophages are known to play a critical role in the survival and chemoresistance of cancer cells. In the case of chronic lymphocytic leukemia (CLL), these tumor-associated macrophages are called Nurse-Like Cells (NLCs) and reside mainly in the lymph nodes, where they are able to protect leukemic B cells (B-CLL) from spontaneous apoptosis and contribute to their chemoresistance, hindering the efficacy of immunotherapy in many patients. NLCs are differentiated from monocytes through cytokines signaling and physical contact with the cancer cells [1], however, the precise mechanisms by which B-CLL cells influence this differentiation are still unknown. We used an in vitro model of leukemia, in which we can closely follow the production of NLCs from monocytes in the presence of leukemic B cells from CLL patients. Building on experimental observations of cancer cells in these cultures of patients' blood, we propose here a two-dimensional agent-based model simulating the monocyte-to-macrophage differentiation and intercellular interactions in the spatial context of this in vitro co-culture of monocytes and cancer B-CLL cells.

Using our time-course measurements of B-CLL cell viability and concentration to optimize the model parameters, we were able to reproduce the experimentally observed dynamics. We further tested the model's predictive power by simulating specific NLC production features in relation to varying measured proportions of monocytes in each patient in the co-cultures. Our results suggest that this model could be made patient-specific using their blood monocytes counts, which is a routinely measured variable. Finally, we performed a sensitivity analysis of the different parameters and suggest a strong role for phagocytosis from monocytes and NLCs to ensure the survival of cancer cells in this in vitro CLL model, especially in the initial phases of the time course. Additionally, we show that the protective anti-apoptotic signals provided by NLCs to the cancer cells are most important towards the later stages of the culture. This finding suggests that monitoring and potentially modulating phagocytosis could play a role in the control of NLCs polarization in CLL and also help understanding tumor-associated macrophages formation even in solid tumors [2].

References

1. Frédéric Boissard et al. Nurse like cells: Chronic lymphocytic leukemia associated macrophages. Leuk. Lymphoma, 56(5):1570- 1572, 2015.
2. Nina Verstraete et al. An Agent-Based Model of Monocyte Differentiation into Tumor-Associated Macrophages in Chronic Lymphocytic Leukemia. https://www.biorxiv.org/content/10.1101/2021.12.17.473137v2

Probing SARS-CoV-2 RNA interactome to unravel post-transcriptional dysregulation associated with COVID-19

Deeya Saha[1], Andreas Zanzoni[1] and Christine Brun[1]

[1] TAGC, INSERM-AMU, Theories and Approaches to Genomic Complexity, Marseille, France

Corresponding Author: andreas.zanzoni@univ-amu.fr; christine-g.brun@inserm.fr

1. Introduction

SARS-CoV-2 greatly remodels host's RNA bound proteome or RBPome [1]. Previous work on RNA-protein interactions have shown that RNA binding proteins (RBPs) regulate functionally related mRNAs encoding proteins involved in the same biological processes [2], the so-called RNA regulons. We have previously studied the pervasiveness of the regulon theory at coding transcriptome level by associating cellular pathways to RBPs. Based on this study [3], we aimed at addressing some fundamental questions on the role of RNA-protein interactions between SARS-CoV-2 genomic RNA (gRNA) and cellular RBPs in the context of infection. We tested whether the sequestration of RBPs could indeed impact the host post-transcriptional regulatory networks. We also sought to determine the range of host biological processes that are likely to be perturbed as a consequence of the interaction between a given RBP and the viral gRNA.

2. Results

Based on our previous approach [3], we built an RBP-pathway association map using a rigorous statistical approach. We observed that RBPs known to be linked to viral infection regulate a significantly higher number of pathways as compared to the others. We gathered experimentally determined SARS CoV-2 gRNA-protein interaction datasets and observed that cellular RBPs that are bound to SARS-CoV-2 gRNA are associated with a large number of pathways as compared to non-gRNA binders. Moreover, it was also observed from our analyses that pathways dysregulated during SARS-CoV-2 infection are more likely to be associated with RBPs that are bound to SARS-CoV-2 gRNA. Thus, our data supports the notion that SARS-CoV-2 gRNA could indeed act as a sponge to sequester different cellular RBPs. As a result of this sequestration, dysregulation of important cellular pathways such as carbohydrate metabolism, could be observed. Our computational strategy also provides some testable hypotheses on G3BP1, a core component of stress granule and SARS-CoV-2 gRNA binder. Our data suggests that G3BP1 is a potential regulator of carbohydrate metabolism. We also hypothesize that sequestration of G3BP1 could have profound effects on various pathways such as neurodegenerative pathways, which can lead to their dysregulation upon SARS-CoV-2 infection.

3. Conclusions

Our study shows the effect of sequestration of cellular RBPs by SARS-CoV-2 gRNA on host post-transcriptional regulatory networks. This underlines that SARS-CoV-2 gRNA-binding proteins are promiscuous and can regulate multiple different pathways. Lastly, we generate testable hypothesis on G3BP1, a known viral RNA binder.

Acknowledgements

The work is funded by RIA H2020-SC1-PHE-CORONAVIRUS-2020 grant to CB (#101003633, RiPCoN project).

References

1. W. Kamel *et al.* Global analysis of protein-RNA interactions in SARS-CoV-2 infected cells reveals key regulators of infection. *bioRxiv*, 2020.
2. J. D. Keene. RNA regulons: Coordination of post-transcriptional events. *Nature Reviews Genetics*, vol. 8, no. 7. 2007.
3. A. Zanzoni, L. Spinelli, D. M. Ribeiro, G. G. Tartaglia, and C. Brun. Post-transcriptional regulatory patterns revealed by protein-RNA interactions. *Scientific Reports*, 2019.

Refined quantitative models for the control of gene expression by IFNα in Primary Sjögren's Syndrome

DianaTrutschel[1], Cheïma Boudjeniba[123], Darragh Duffy[4], Jacques Eric Gottenberg[5] and Benno Schwikowski[1]

[1] Institut Pasteur, Computational Systems Biomedicine lab, 25-28 Rue du Dr Roux, 75015 Paris, France
[2] Servier, 50 Rue Carnot, 92150, Suresnes
[3] Université Paris Cité, Laboratoire MAP5 UMR 8145, Paris, France
[4] Institut Pasteur, Translational Immunology lab, 25-28 Rue du Dr Roux, 75015 Paris, France
[5] Les Hôpitaux Universitaires de Strasbourg, Rhumatologie, 1 place de l'hôpital, Strasbourg, France

Corresponding author: diana.trutschel@pasteur.fr

1 Introduction

Primary Sjögren's Syndrome (pSS) is a *systemic chronic autoimmune disorder* with common symptoms fatigue, dryness and pain. To date, the *treatment of pSS remains symptomatic*, although research improves the understanding of the disease [1]. The *type I interferone (IFNα) system has a pivotal role* in the disease process and is associated with clinical and immunological phenotype. However, the impact of IFNα on different aspects of the disease remains unresolved.

2 Objective

We aimed to improve our understanding of the clinical correlates of varying IFNα levels through quantitative modeling that relate quantitative IFNα protein measurements with gene expression. Since IFNα levels are known to be associated with autoantibodies, like anti-SSA and anti-SSB, presentation, as well as disease activity, our approach was to explain and quantify their *different type of interplay* to effect gene expression. Different gene signatures indicate then gene groups with different interaction to genes.

3 Method

Data from patients with pSS from the multi-center prospective clinical cohort ASSESS (ClinicalTrials.gov ID: NCT03040583) were analyzed, whereby transcriptomic data (GEO: GSEA 140161) as well as parameters, like IFNα and levels of the autoantibodies SSA and SSB, were available. We tested 6 different models for their ability to capture different types of interplay and to identify the best fitting model for each gene. We employed *linear mixed regression models*, adjusting for the effect of age including as a fixed factor and for hospital center effects as a random factor.

4 Results

Three main gene signatures were identified, whereby the analysis showed that most of the genes are *related to both* IFNα and autoantibody level instead of only one of them. Furthermore, in general the gene expression increases with increasing IFNα blood concentration, but the increase is higher the more autoantibodies are present. This effect of autoantibodies to the increase of gene expression by increasing IFNα was found to be constant for those genes related to both factors. Moreover, the impact of the second antibody (SSB) on gene expression was – in a *surprisingly consistent* manner – found to be *three times as strong* as the impact of the first antibody (SSA).

5 Discussion

Our models reveal different classes of genes whose expression is related in different ways to IFNα and autoantibody levels in Primary Sjögren's Syndrome. The quantitative nature of these models opens the door for *further refinements*, such as the precise characterization of the effect of other known major players controlling gene expression, such as the rheumatoid factor.

References

[1] G. Nocturne and X. Mariette. Advances in understanding the pathogenesis of primary sjögren's syndrome. *Nat Rev Rheumatol*, 9, 2013.

Galaxy-SynBioCAD: tools and automated pipelines for Synthetic Biology Design and Metabolic Engineering

Thomas Duigou[1]*, Joan Hérisson[2]*, Melchior du Lac[1,3], Kenza Bazi Kabbaj[1], Mahnaz Sabeti Azad[1], Gizem Buldum[4], Olivier Telle[1], Yorgo El Moubayed[1], Pablo Carbonell[5,6], Neil Swainston[5,7], Valentin Zulkower[8], Manish Kushwaha[1], Geoff S. Baldwin[4] and Jean-Loup Faulon[1,2,5]

[1] Université Paris-Saclay, INRAE, AgroParisTech, Micalis Institute, 78352 Jouy-en-Josas, France
[2] Genomics Metabolics, Genoscope, François Jacob Institute, CEA, CNRS, Évry University, Paris-Saclay University, 91057 Evry, France
[3] Amyris Inc, Emeryville, CA, 94608-2405, US
[4] Imperial College Centre for Synthetic Biology, Department of Life Sciences, Imperial College, London, SW7 2AZ, UK
[5] Manchester Institute of Biotechnology, SYNBIOCHEM center, School of Chemistry, University of Manchester, Manchester M1 7DN, UK
[6] Institute of Industrial Control Systems and Computing (ai2), Universitat Politècnica de València, 46022 Valencia, Spain
[7] Institute of Systems, Molecular and Integrative Biology, University of Liverpool, Liverpool L69 7ZB, UK
[8] Edinburgh Genome Foundry, SynthSys, School of Biological Sciences, University of Edinburgh, EH93BF Edinburgh, UK

Corresponding author: jean-loup.faulon@inrae.fr

Abstract

There's a substantial number of tools released around Synthetic Biology and Metabolic Engineering-related questions and needs. This population of tools is difficult to comprehend and use together, main reasons being complexity and interoperability issues. Indeed, a high level of expertise could be required for installing codes, and execution for real life use cases could be computationally resource demanding. Plus some tools, although complementary, use different inputs and outputs which prevent easy chaining.

The SynBioCAD-Galaxy portal is a growing toolshed for synthetic biology, metabolic engineering, and industrial biotechnology[1]. The tools and workflows currently shared on the portal enable one to build libraries of strains producing desired chemical targets covering an end-to-end metabolic pathway design and engineering process: from the selection of strains and targets, the design of DNA parts to be assembled, to the generation of scripts driving liquid handlers for plasmid assembly and strain transformations.

Tools are made available on GitHub, anaconda.org and the Galaxy Tool Shed, opening to the greatest number access and utilization throughout the SynBio community, and significant effort has been granted for adopting FAIR principles. As a community effort helped by funded projects, the scope covered by tools is expected to expand over time.

The poster will give an overview of the SynBioCAD-Galaxy portal in the context of prediction and construction of *E. coli* lycopene-producing pathways. The poster will open the discussion around good practices guiding releases of tools through continuous integration.

A – lightweight – testing instance of SynBioCAD-Galaxy is available at https://galaxy-synbiocad.org.

References

[1] Joan Hérisson, Thomas Duigou, Melchior du Lac, Kenza Bazi Kabbaj, Mahnaz Azad Sabeti, Gizem Buldum, Olivier Telle, Yorgo El Moubayed, Pablo Carbonell, Neil Swainston, Valentin Zulkower, Manish Kushwaha, Geoff S. Baldwin, and Jean-Loup Faulon. Galaxy-synbiocad: Automated pipeline for synthetic biology design and engineering. *bioRxiv*, 2022.

RFLOMICS : R package and Shiny interface for Integrative analysis of omics data

Nadia BESSOLTANE[1], Christine PAYASANT-LE-ROUX[2], Gwendal CUEFF[1], Audrey HULOT[1], Delphine CHARIF[1]

1Institut Jean-Pierre Bourgin (IJPB), Université Paris-Saclay, INRAe, AgroParisTech, 78000, Versailles, France

2 Institute of Plant Sciences Paris-Saclay (IPS2), Université Paris-Saclay, CNRS, INRAE, Univ Evry, 91190, Gif-sur-Yvette, France

Corresponding Author: nadia.bessoltane@inrae.fr and delphine.charif@inrae.fr

The acquisition of multi-omics data in the context of complex experimental design is a widely used practice to identify entities and decipher the biological processes they are involved. The investigation of each omics layer is a good first step to explore and extract relevant biological variability. The statistical integration could then be restrained to pertinent omics levels and features. Such analysis of heterogeneous data remains a technical challenge with the needs of expertise methods and parameters to take into account data specificity. Furthermore, applying different statistical methods from several tools is also a technical challenge in term of data management. In this context, we developed RFLOMICS, an R package with a shiny interface, to ensure the reproducibility of analysis, with a guided and comprehensive analysis and visualization of data in a framework which can manage several omics-data and analysis results.

RFLOMICS currently supports up to three types of omics (RNAseq, proteomics, and metabolomics), and can deal with multi-factorial experiments (up to 3 biological factors). It includes methods chosen based on expert feedback [1,2]. This application is divided into three key steps. The first step allows the user to import the experimental design file and abundance matrix for each dataset (read counts for RNA-Seq, signal intensity for metabolomics and proteomics), and set up the statistical model and contrasts associated to the biological issues [1]. The second step is to perform a full analysis for each dataset, which includes : i- quality control to check for batch effects or identify outlier samples that can be removed, ii- filtering and normalization of RNA-Seq data, or transformation of prot/meta data, iii- differential expression analysis using edgeR [1] for RNA-Seq and limma [3] for prot/meta data, iv- co-expression analysis using coseq [1], and finally, v- functional enrichment analysis [1]. The third step is to integrate selected omics layers using the unsupervised methods proposed by MOFA [4]. All the results as well as the raw data, and all information necessary for reproducibility of analysis are managed and stored thanks to the MultiAssayExperiment object [5]. An HTML report can be generated, summarizing all analysis steps, using rmarkdown R package.

RFLOMICS provides the same framework that allows the user to perform the analysis of multi-omics project from A to Z, taking into account the complexity of the design. It guarantees the relevance of the used methods, and ensures the reproducibility of the analysis. The interface offers an interesting flexibility between the visualization of the results and the data manipulation (filtering, parameter setting). Future development will include the implementation of supervised integration methods, and a docker image to facilitate deployment.

Acknowledgements

This project is part of CATI SySmics [2]. We thank ML. Martin-Magniette, V. Brunaud, G. Rigaill (IPS2) for helpful discussion and feedback. We also thank L. Rajou (IJPB) for given access to Ecoseed project data.

References

1. Lambert, I, *et al*. (2020). “DiCoExpress: a tool to process multifactorial RNAseq experiments from quality controls to co-expression analysis through differential analysis based on contrasts inside GLM models”. Plant Methods.
2. https://sysmics.cati.inrae.fr/
3. Efstathiou G, *et al.* (2017). ProteoSign: an end-user online differential proteomics statistical analysis platform. Nucleic Acids Res.
4. Argelaguet R, *et al.* (2018). “Multi-Omics Factor Analysis a framework for unsupervised integration of multi-omics data sets.”Molecular Systems Biology.
5. Ramos M, *et al.* (2017). Software For The Integration Of Multi-Omics Experiments In Bioconductor. Cancer Research.

Impact environnementaux et sociaux du numérique

David Benaben[1,2]
[1] UMR Biologie du Fruit et Pathologie, Université de Bordeaux, INRAE, F-33140 Villenave d'Ornon, France
[2] Bordeaux Metabolome, MetaboHUB, PHENOME-EMPHASIS, F-33140 Villenave d'Ornon, France

Auteur référent: david.benaben@inrae.fr

L'urgence climatique, l'empreinte environnementale de l'homme grandissante ou le dépassement de limites planétaires nous oblige à nous interroger sur nos pratiques et nos impacts. Notamment ceux du numérique, au cœur des activités d'un bioinformaticien.

Si ce n'est pas le poste le plus important en terme d'émission de gaz à effet de serre (GES) parmi toutes nos activités (transport, alimentation, bâtiment, etc), le numérique représenterait quand même 2 à 4% des GES [1]. En France, ces impacts sont principalement liés à la fabrication (78%) et aux terminaux utilisateurs (79%) [2]. Mais le numérique, c'est aussi des minerais (600Kg de matières premières mobilisées pour la fabrication d'un ordinateur). Cette exploitation minière, industrie parmi les plus polluantes au monde, provoque localement des bouleversements sociaux-économiques, environnementaux, sanitaires et parfois des violations de droits humains. C'est aussi de l'eau (0,2% eau douce mondiale) pour l'extraction de métaux ou la fabrication de composants électroniques. C'est encore des déchets pas toujours collectés et peu recyclés.

Et ce secteur est en croissance (en contradiction avec les objectifs de réduction globale) avec des effets rebonds difficiles à évaluer.

Les traitements réalisés en bioinformatique (alignement, assemblage, docking, etc.) s'appuient sur les équipements informatiques. On peut alors essayer d'évaluer ces impacts [3], de même que certains centres de calcul ont estimé leur empreinte carbone. On notera notamment le GRICAD et la plateforme bioinformatique GenoToul qui évaluent l'heure par cœur de calcul autour de 5g eqCO2 [4,5]. A cette empreinte du numérique on peut ajouter le stockage des données, les transferts, les visioconférences, nos terminaux utilisateurs, etc. Ces impacts peuvent être mitigés en rallongeant la durée de vie des équipements, en mutualisant, en améliorant l'efficacité énergétique, etc.

Finalement, si on estime nos projets nécessaires, il faut allonger autant que possible la durée de vie des équipements, réduire leur nombre et tendre vers une certaine sobriété de nos usages.

Remerciements

Merci au GDS EcoInfo pour les formations et les nombreux travaux, ainsi qu'à l'Institut Français de Bioinformatique (IFB) (ANR-11-INBS-0013) pour l'accompagnement.

Références

[1] Charlotte Freitag, Mike Berners-Lee, Kelly Widdicks, Bran Knowles, Gordon S. Blair, and Adrian Friday. The real climate and transformative impact of ict : A critique of estimates, trends, and regulations. *Patterns*, 2(9) :100340, 2021.

[2] ADEME. Le numérique : quels impacts environnementaux ?, 2022.

[3] Jason Grealey, Loic Lannelongue, Woei-Yuh Saw, Jonathan Marten, Guillaume Meric, Sergio Ruiz-Carmona, and Michael Inouye. The carbon footprint of bioinformatics. *BioRxiv*, 2021.

[4] Francoise Berthoud, Bruno Bzeznik, Nicolas Gibelin, Myriam Laurens, Cyrille Bonamy, Maxence Morel, and Xavier Schwindenhammer. Estimation de l'empreinte carbone d'une heure.coeur de calcul. Research report, UGA - Université Grenoble Alpes ; CNRS ; INP Grenoble ; INRIA, April 2020.

[5] Genotoul Bioinfo. Newsletter #35, special carbon footprint issue, 2021.

EMERGEN-DB : The French database for SARS-CoV-2 genomic surveillance and research

Imane MESSAK[1], Anliat MOHAMED[1], Chiara ANTOINAT[1], Arthur LE BARS[1,3], Arianna TONAZZOLLI[1], Benjamin DEMAILLE[1,5], Olivier SAND[1], François GERBES[1,3], Thomas ROSNET[1,4], Laurent BOURI[1,5], Julien SEILER[5,1], Nicole CHARRIÈRE[1], Christophe ANTONIEWSKI[6], Anne BOZORGAN[7], Javier CASTRO ALVAREZ[7], Jeanne SUDOUR[7], Yann LE STRAT[7], Bruno COIGNARD[7], Abdelkader Amzert[8], Nebras Gharbi[8], Franck Lethimonier[8], Hélène CHIAPELLO[1,2], Naira NAOUAR[6], Claudine MEDIGUE[1,9], Gildas LE CORGUILLE[3,1], David SALGADO[1,10], Jacques VAN HELDEN[1,4] and Thomas DENECKER[1]

[1] CNRS, Institut Français de Bioinformatique, IFB-core, UMS 3601, Évry, France

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

[3] Sorbonne Université, CNRS, FR2424, ABiMS, Station Biologique, 29680, Roscoff, France

[4] Aix-Marseille Univ, Inserm, laboratoire Theory and approaches of genome complexity (TAGC), Marseille, France

[5] CNRS UMR7104, Inserm U1258, Université de Strasbourg, Institut de Génétique et de Biologie Moléculaire et Cellulaire, Illkirch, France

[6] Sorbonne Université, CNRS FR3631, Inserm US037, Institut de Biologie Paris Seine (IBPS), ARTbio Bioinformatics Analysis Facility, Paris, France

[7] Santé Publique France, 12, rue du Val d'Osne 94 415 Saint-Maurice Cedex

[8] Inserm, Institut national de la santé et de la recherche médicale, 101 rue de Tolbiac 75013 Paris.

[9] UMR 8030, CNRS, Université Evry-Val-d'Essonne, CEA, Institut de Biologie François Jacob - Genoscope, Laboratoire d'Analyses Bioinformatiques pour la Génomique et le Métabolisme, Evry, France

[10] Aix Marseille Univ, INSERM, MMG, 13005, Marseille, France

Corresponding Author: imane.messak@france-bioinformatique.fr

In January 2021, the French Ministries of Health (MSS) and Research (MESRI) launched EMERGEN, a national plan for SARS-CoV-2 genomic surveillance, which aims at monitoring the evolution of the COVID-19 in France, detecting new variants and supporting the integration of viral genomic data and health data for both surveillance and research. We present two components of the EMERGEN-Bioinfo digital platform developed by the IFB (Institut Français de Bioinformatique):

(a) **EMERGEN-DB**, the database that collects and manages non-sensitive metadata (sample collection, sequencing method, …) and consensus SARS-CoV-2 genomic sequences produced by the 55 sequencing platforms of the consortium. Developed under Django, it is equipped with both user-friendly and application programmatic interfaces (API) that currently offer over 121 entry points, enabling users to upload their data and query the database manually or in batch. EMERGEN-DB also offers numerous tools to facilitate real-time monitoring of variant evolution and expansion throughout France (alerts, etc.), data export via the APIs, data brokering services to curate and manage the metadata before submitting them to the international repositories such as GISAID and EBI-ENA, and finally data exploration and visualization (summary tables, figures, maps, …) through the interactive pages (activity, sampling, …).

(b) **Rtools4emergen,** an R package enabling to query EMERGEN-DB, to visualize data and to automatically generate reports.

To date, EMERGEN-DB collects metadata from ~74 sequencing platforms and has gathered more than 520513 records, of which 120186 were submitted to GISAID via our brokering tool.

Keywords SARS-CoV-2; COVID-19; EMERGEN-Bioinfo; EMERGEN-DB; genomic surveillance;

Omics Data Analysis Facilities in a Biomedical Research Institute

Justine GUÉGAN[1], Beáta GYÖRGY[1], Thomas GAREAU[1], Emeline CHERCHAME[1], Riwan BRILLET[1], Corentin RAOUX[1] and Violetta ZUJOVIC[1]

[1] Data Analysis Core, Institut du Cerveau, Inserm, CNRS, Sorbonne Université, Paris, France

Corresponding Author: beata.gyorgy@icm-institute.org

Keywords: bioinformatics, NGS, shiny, biomedical, core facility

The Data Analysis Core – DAC - facility is part of Paris Brain Institute (ICM), which is dedicated to basic and clinical neuroscience research; it develops and makes available software solutions and methodological expertise in three domains: data management (curation, standardization, structuration, integration); high throughput genetics and omics data processing (in particular from NGS data); basic and advanced biostatistics, especially integration of multimodal data (namely clinical, omics and imaging data).

As part of the omics data analysis activity, a dedicated team of 6 persons within the platform assists scientific and clinical teams from the design of their study up to data processing, analysis and interpretation. This support consists of three complementary services: the building and operation of specialized pipelines to compute the raw data; the development and deployment of graphical tools to help in the interpretation of the results; a personalized assistance to biologists to go deeper in their scientific questions.

Software pipelines were built around three technologies: Snakemake [1], a workflow manager that makes pipelines scalable by enabling their parallelization; Conda [2], a package manager used to make the installation of the pipelines and their dependencies automatic, and Illumina DRAGEN (https://www.illumina.com/products/by-type/informatics-products/dragen-bio-it-platform.html), a fieldprogrammable gate array technology (FPGA) that provides hardwareaccelerated implementations of common genomic analysis algorithms. Pipelines were developed for the following types of (epi)genomics studies: gene panel, whole-exome (WES) and whole-genome (WGS) sequencing (SNPs, CNVs, expansions, rare variants), bulk RNA-seq (differential gene expression, fusion transcript detection, small and long non-coding RNA), single-cell/nuclei RNAseq (CITE-seq, multiome), bisulfite-seq (methylation profile) [3], ATAC-seq (chromatin accessibility), and ChIP-seq (protein binding). Currently, two pipelines are in development for the analysis of spatial transcriptomics data from NanoString, and for the analysis of long-reads sequencing from Nanopore.

Shiny/R graphical applications were developed to make data available to end-users in an intuitive and interactive way. Dedicated tools are thus proposed to explore WGS/WES data, transcriptomics data from bulk RNA-seq experiments, and a dedicated module for the analysis of single-cell RNAseq. In addition, a recent development was performed to build DEJAVU, a MongoDB database of variants identified in the frame of ICM studies, together with a graphical user interface. All those interfaces are managed in one tool named QUBY, with Shinyproxy, an open-source system that makes it possible to deploy dockerized applications, with a built-in functionality for Keycloak authentication and authorization, which makes securing Shiny traffic (over TLS) a breeze and has no limits on concurrent usage of a Shiny app.

Finally, ad hoc expertise is proposed as a follow-up of every project. The bioinformaticians of the platform dialog with biologists and clinicians to understand their scientific questions, and extract relevant information from experimental results. This activity consists in guiding scientists from and outside ICM in the use of softwares and methods, and developing scripts to carry out specific data processing steps; this is tightly linked to the biostatistics component of DAC. Co-authored publications are a frequent outcome of such projects.

References

1. Köster J and Rahmann S. Snakemake – a scalable bioinformatics workflow engine. *Bioinformatics*, 28:2520-2522, 2012.
2. Anaconda Software Distribution. Computer software. Vers. 2-2.4.0, Nov. 2016. Web. https://anaconda.com

The IFB Catalogue

Bryan Brancotte[1], Hippolyte Kengni[2], Thomas Rosnet[2,3], Laurent Bouri[2,6], Jon Ison[2], Olivier Sand[2], Hélène Chiapello[4,2], Alban Gaignard[5], Sylvain Milanesi[2], Jacques van Helden[2,3]and Hervé Ménager[1,2]

[1] Institut Pasteur, Université Paris Cité, Bioinformatics and Biostatistics Hub, F-75015 Paris, France
[2] CNRS, Institut Français de Bioinformatique, IFB-core, UAR 3601, Evry, France
[3] Aix-Marseille Univ, INSERM, Lab. Theory and Approaches of Genome Complexity (TAGC), Marseille, France
[4] INRAE, Université Paris-Saclay, MaIAGE, Jouy-en-Josas, France
[5] Nantes Université, CNRS, INSERM, l'institut du thorax, F-44000 Nantes, France
[6] Université de Strasbourg, CNRS, INSERM, Institut de Génétique et de Biologie Moléculaire et Cellulaire, IGBMC, Illkirch, France

Corresponding author: herve.menager@pasteur.fr

The IFB Catalogue is a centralized database developed as part of the IFB Distributed national environment of services in Bioinformatics. Its aim is to ensure the visibility and accessibility of the Bioinformatics resources provided and maintained by the french community, whether these are research labs or service platforms, in a structured and open database that guarantees their FAIRness. Such resources can be software tools, databases, computing resources, individual expertises, platforms, trainings and training materials. This catalogue stores the metadata describing their properties (*e.g.* the licence of a software tool) and linking them (*e.g.* the publication of a database by a given team).

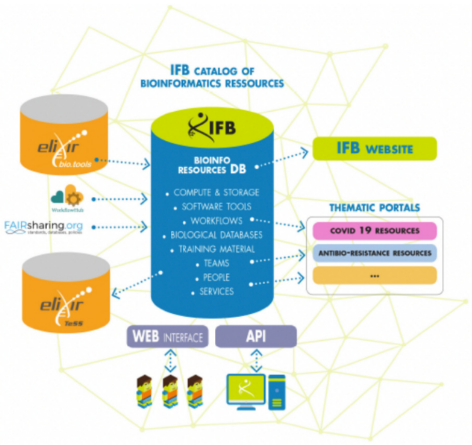


Fig. 1. Architecture of the IFB Catalogue

The primary role of the Catalogue is to help users of Bioinformatics services. Through the integration of the data in the IFB website (https://www.france-bioinformatique.fr), it provides an overview of the various resources provided by the community, and supports the needs of end-users (who can perform a given kind of analysis? where is a tool available?). It is also synchronized with related international catalogues (*e.g.* ELIXIR bio.tools[1] , TeSS[2]), and promotes the reusability of the data through the publication of Bioschemas[3] markup and REST APIs as well as the use of the EDAM ontology[4].

The backend server is available at https://catalogue.france-bioinformatique.fr/. The code of the Catalogue database backend is openly available on https://github.com/IFB-ElixirFr/ifbcat.

Acknowledgements

The authors would like to thank the participants representing the various IFB platforms for their patience, understanding, and valuable improvement suggestions during the successive Catalogue curation sessions (a.k.a. *Catathons*). Fig. 1 was designed by multiple members of Institut Français de Bioinformatique IFB/ELIXIR-FR. This service would also not be available without the help of Nicole Charrière, our systems administrator.

References

[1] Jon Ison, Hans Ienasescu, Piotr Chmura, Emil Rydza, Hervé Ménager, Matúš Kalaš, Veit Schwämmle, Björn Grüning, Niall Beard, Rodrigo Lopez, et al. The bio. tools registry of software tools and data resources for the life sciences. *Genome biology*, 20(1):1–4, 2019.

[2] Niall Beard, Finn Bacall, Aleksandra Nenadic, Milo Thurston, Carole A Goble, Susanna-Assunta Sansone, and Teresa K Attwood. Tess: a platform for discovering life-science training opportunities. *Bioinformatics*, 36(10):3290–3291, 2020.

[3] Alasdair JG Gray, Carole A Goble, and Rafael Jimenez. Bioschemas: from potato salad to protein annotation. In *International Semantic Web Conference (Posters, Demos & Industry Tracks)*, 2017.

[4] Jon Ison, Matúš Kalaš, Inge Jonassen, Dan Bolser, Mahmut Uludag, Hamish McWilliam, James Malone, Rodrigo Lopez, Steve Pettifer, and Peter Rice. Edam: an ontology of bioinformatics operations, types of data and identifiers, topics and formats. *Bioinformatics*, 29(10):1325–1332, 2013.

IFB training activities and resources

Lucie KHAMVONGSA-CHARBONNIER [1], Yousra MAHMAH[1], Jacques VAN HELDEN[1], Olivier SAND[1] and Hélène CHIAPELLO[1,2]

[1] CNRS, Institut Français de Bioinformatique, IFB-core, UMS 3601, Évry, France
[2] Université Paris-Saclay, INRAE, MaIAGE, 78350 Jouy-en-Josas, France

Corresponding Author: lucie.khamvongsa-charbonnier@france-bioinformatique.fr

The French Institute of Bioinformatics (IFB) is the National Bioinformatics Infrastructure that provides support, deploys services, organizes training and carries out innovative developments for the life science communities. According to a 2019 IFB survey targeting life scientists and bioinformaticians, almost all teams and units express the need for training. The most requested skills are related to NGS analysis, biostatistics and data analysis (including machine learning and AI), and bioinformatics skills, especially related to the recent field of integrative bioinformatics. The survey also highlights an important need for training in the fundamentals of Open Science and FAIR principles applied to bioinformatics activities.
In order to face this important demand, IFB training activities are organized at 3 levels :
- at the regional level with the training activities carried out by the IFB platforms and teams,
- at the national level with actions (co)-coordinated by the IFB,
- at the European level through the ELIXIR network.
In this poster we will report a summary of the IFB training activities over the last three years. We will also present recent IFB training actions that have been designed to enhance bioinformatics skills at the national level and fill gaps in the existing french training activity landscape. We will focus on two new courses designed in 2020 and 2021: (i) The FAIR bioinfo training, which aims to make bioinformatics analyses more reproducible by implementation of the FAIR principles in a bioinformatics analysis or development project and (ii) The FAIR data training, which is dedicated to the fundamental aspects of Open Data, including legal aspects, practical DMP sessions and metadata issues in the context of omics data and bioinformatics. We will also present future projects for the development of new training resources, e-learning material and ongoing training actions related to integrative bioinformatics skill development.

Training events, Training materials, Open Science, FAIR, e-learning

The Assemblathon of the UAR 2AD, Data Acquisition and Analysis for Natural History

Marie CARIOU[1], Jawad ABDELKRIM[1] and Julien MOZZICONACCI[1]

[1] Unité d'Appui et de Recherche 2700 Acquisition et Analyse de Données pour l'Histoire Naturelle, 43 rue Cuvier - étage 1 - CP26, 75005 , Paris, France

Corresponding Author: marie.cariou@mnhn.fr

Recent technological developments have enabled telomere-to-telomere assembly of genomes for numerous non model species [1]–[3]. Among these progresses, the continual improvement of sequencing techniques, in particular regarding the sequencing of long fragments plays an important role, as well as the application to scaffolding of methods to acquire long-range contiguity information such as conformation capture [4]. Also, assembly algorithms and pipelines, including important preprocessing and polishing steps, are developed to tackle issues that frequently hamper assembly of non-standard genomes (e.g. with high ploidy or heterozygosity and repeated regions) [5]. However, genome assembly remains a difficult task. Empirical optimisation are still often required for each dataset, making full automation difficult to achieve.

The National Museum of Natural History of Paris (MNHN) stands out for the diversity of its research models. Many of its research projects ambition to build genomic data from non-standard organisms and/or DNA sources. The Data Analysis Service (SAD) belongs to the UAR 2700 2AD, Data Acquisition and Analysis for Natural History. Its aim is to provide support to researchers regarding their data analysis notably regarding NGS data. Both to fulfil this ambition and to take advantage of this multiplicity of projects involving genomic data generation, we are organizing an Assemblathon, which aims are:

(1) to create and maintain flexible assembly pipelines, adapted to a high diversity of organisms, sequencing techniques and objectives.

(2) to centralize experiences, original approaches and good practices regarding assembly of non-model organisms.

(3) Eventually, this framework will favor the emergence of new tools and guidelines useful for any new genome assembly project.

This poster brushes over the initial set-up of our assembly pipeline, preliminary tests as well as an overview of the panel of projects currently participating to the Assemblathon.

Acknowledgements

We thank Bertrand Bed'Hom, Amélie Chimènes, Evelyne Duvernois-Berthet, Marc Eléaume, Sarah Farhat, Bernardo Ferreira Dos Santos and Benjamin Marie for their participation to the Assemblathon of the UAR 2AD.

References

[1] N. Guiglielmoni, R. Rivera-Vicéns, R. Koszul, et J.-F. Flot, « A Deep Dive into Genome Assemblies of Non-vertebrate Animals », LIFE SCIENCES, preprint, nov. 2021. doi: 10.20944/preprints202111.0170.v1.

[2] P. Simion *et al.*, « Chromosome-level genome assembly reveals homologous chromosomes and recombination in asexual rotifer *Adineta vaga* », *Sci. Adv.*, vol. 7, nº 41, p. eabg4216, oct. 2021, doi: 10.1126/sciadv.abg4216.

[3] S. Farhat *et al.*, « Comparative analysis of the Mercenaria mercenaria genome provides insights into the diversity of transposable elements and immune molecules in bivalve mollusks », *BMC Genomics*, vol. 23, nº 1, p. 192, déc. 2022, doi: 10.1186/s12864-021-08262-1.

[4] L. Baudry *et al.*, « instaGRAAL: chromosome-level quality scaffolding of genomes using a proximity ligation-based scaffolder », *Genome Biol.*, vol. 21, nº 1, p. 148, déc. 2020, doi: 10.1186/s13059-020-02041-z.

[5] A. Rhie *et al.*, « Towards complete and error-free genome assemblies of all vertebrate species », *Nature*, vol. 592, nº 7856, p. 737-746, avr. 2021, doi: 10.1038/s41586-021-03451-0.

Team efforts of the Bioinformatics and Biostatistics Hub of Institut Pasteur in response to the COVID-19 pandemic

Bioinformatics and Biostatistics Hub [1]

1 Institut Pasteur, Université Paris Cité, Bioinformatics and Biostatistics Hub, F-75015 Paris, France

Corresponding Author: julien.fumey@pasteur.fr

The COVID-19 pandemic due to the SARS-CoV-2 started late 2019 in China [1]. The first case was detected in France late January 2020. Institut Pasteur isolated the virus and sequenced it. Rapidly, the pandemic became a major health and economical issue with many populations under lockdown throughout the world.

The Bioinformatic and Biostatistics Hub at Institut Pasteur became an asset for the fight against the virus at Institut Pasteur. We are involved in different aspects of the epidemiological monitoring of COVID-19 and SARS-CoV-2 related research.

We work in close relationship with the National Reference Center for Respiratory Viruses. In this collaboration, the main task of the Hub was the development of an automatic pipeline to assemble the sequences generated using the Illumina platform at the P2M sequencing platform. The pipeline validates if the consensus sequences it generates pass the acceptance criteria to be integrated in the public database GISAID [2] dedicated to Flu and COVID. It also uses Nextclade [3] and Pangolin [4] to identify the clade of each sample. For each sequence with either borderline QC indicators or unusual mutational events, Hub engineers review thoroughly the consensus sequence and the aligned sequencing reads before deciding to submit the sequence to public repositories. Once the sequences are validated, they are submitted to GISAID together with metadata describing the samples. The COVID pipeline and curation procedure are now also adapted to track other viruses like Influenza.

Another aspect of our support efforts consist of the development of tools to facilitate the epidemiological monitoring of the pandemic, either by the CNR (website presenting statistics about the french sequences published on public repositories) or by the general population with a website “modélisation COVID” developed to share the outcomes of the modeling of the dynamic of the pandemic done at Institut Pasteur.

Our contribution to research effort covers a wide variety of topic such as the search for natural relative of SARS-CoV-2, the characterization of the immune response to infection, inferring specific variant phylodynamics or studies of questions relevant to decision makers in public health (e.g: whether a booster vaccine protects against Omicron variant).

The Hub has also helped the GISAID consortium with the curation of the data that were submitted to the database, developed a curation tool to automate this curation process and an advanced efficient dedicated alignment tool COVID-Align.

In this poster we will present these different parts of the involvement of the Bioinformatics and Biostatistics Hub in the COVID-19 answer at Institut Pasteur and how this involvement can be translated to track other viruses like Influenza virus.

References

[1] Wu F., et al. (2020) A new coronavirus associated with human respiratory disease in China. Nature 579:265–269.

[2] Shu Y. and McCauley J. (2017) GISAID: from vision to reality. EuroSurveillance, 22(13)

[3] Aksamentov, et al. (2021). Nextclade: clade assignment, mutation calling and quality control for viral genomes. Journal of Open Source Software, 6(67):3773

[4] O’Toole A., et al. (2021) Assignment of epidemiological lineages in an emerging pandemic using the pangolin tool, Virus Evolution, 7(2)

BioInformatics and Genomics platform at Institut Sophia Agrobiotech

Martine DA ROCHA[1], Arthur PÉRÉ[1], Etienne G.J. DANCHIN[1] and Corinne RANCUREL[1]

[1] Institut Sophia Agrobiotech, INRAE, Université Côte d'Azur, CNRS, 400 route des Chappes BP 167, F-06903 Sophia Antipolis Cedex, France

Corresponding Author: martine.da-rocha@inrae.fr, corinne.rancurel@inrae.fr

The BioInformatics and Genomics (BIG) platform of Institut Sophia Agrobiotech (ISA: INRAE - CNRS - Univ. Côte d'Azur) offers expertise in bioinformatics and solutions for processing, integrating, analyzing and visualizing multi-omics data in the field of plant health and protection. The BIG platform is part of PlantBios (Biocontrol and Plant Biostimulation, Facilities and Expertise), labeled as a collective scientific infrastructure by INRAE. PlantBIOs offers equipment and expertise for studies ranging from gene level to the whole agroecosystem scale with analytical tools (Imagery and Microscopy, Biochemistry and Mass Spectrometry, Bioinformatics and Genomics), experimental tools, and collections of rare biological resources.

Since the end of 2020, BIG has been an IFB contributing platform. The core of the BIG platform is composed of three bioinformatics engineers: Martine Da Rocha and Arthur Péré (from INRAE), with Corinne Rancurel (from CNRS) as operational manager . The core is complemented by a scientific advisor: Etienne Danchin (INRAE senior scientist).

BIG has a main expertise in comparative genomics, transcriptomics, and molecular evolution. More recently, BIG has been involved in epigenomics, small RNA as well as metagenomics studies. The tools and resources produced by BIG are made available to the scientific community (website, forge and integrative portals) and can address similar problems encountered in other research areas. For instance, the Alienness tool, which allows rapid detection of candidate horizontal gene transfers in genomes has been used 1446times by 353 different users and the corresponding paper [1] (Rancurel et al. 2017) has already been cited 29times, since its launch in August 2017.

In addition to methodological developments, the platform offers support and training for biologists in the use of bioinformatics tools and pipelines, including the one developed by BIG itself.

The BIG platform is open for collaboration and can be contacted at the following e-mail address: big.plantbios@inrae.fr

This web page summarizes the activities and organization of the BIG platform: https://www6.paca.inrae.fr/institut-sophia-agrobiotech/Infrastructure-PlantBIOs/Equipements-Ressources-biologiques-et-Expertises/Plateau-de-bioinformatique or http://tinyurl.com/y9qkho4v

References

1. Rancurel C, Legrand L, Danchin EGJ. Alienness: Rapid Detection of Candidate Horizontal Gene Transfers across the Tree of Life. Genes (Basel). 2017 Sep 29;8(10):248. doi: 10.3390/genes8100248. PMID: 28961181; PMCID: PMC5664098.

Montpellier GenomiX (MGX) : next-generation sequencing and data analysis service and expertise

Mathilde ESTEVEZ-VILLAR, Simon GEORGE, Anne-Alicia GONZALEZ, Elise GUERET, Hugues PARRINELLO, Dany SEVERAC, Anaïs LOUIS, Xavier MIALHE, Stéphanie RIALLE and Laurent JOURNOT

Montpellier GenomiX, 141 rue de la cardonille, 34090, Montpellier, France

Corresponding Author: xavier.mialhe@mgx.cnrs.fr

https://www.mgx.cnrs.fr/

Abstract*:*
The MGX (Montpellier GenomiX) is an NFX 50-900 certified facility which offers, since 2008, next-generation sequencing services, as well as bioinfomatics and biostatistics analysis of the produced data. The facility is accessible to both academic and industry/biotech scientists. Our expertise is drawing on many years of experience, as much in molecular biology as in bioinformatics.

Keywords : Sequencing, analysis, service, libraries, expertise

Our facility offers a wide range of applications on Novaseq 6000, including whole-genome sequencing, exome and targeted sequencing, RNA-seq, small RNA-seq, epigenetics (ChIP-seq, HiC, Whole Genome and Reduced Representation Bisulfite Sequencing, …), population genomics with RAD-seq, etc. We an integrated service from library construction to bioinformatics analysis. Bioinformatics analyses include quality control, alignment of sequences to a reference genome or transcriptome, statistical and functional analyses. A typical project starts with a launch meeting to define the aims of the experiment, the experimental design, and the analysis tools to be used. Throughout the project, a project management web application provides an easy and flexible way to store and retrieve information, and to communicate with customers.

Using a Chromium device from 10X genomics, we offer single cell gene expression, as well as ATAC and Multiplex analysis. The Chromium is a droplet-based method, which allows the characterization of hundreds to thousand of cells in a single experiment. The 3’ mRNA quantification of gene expression is used to identify cell populations in heterogenous or complex samples. The analysis can be done using the Cell Ranger solution available from 10x Genomics. We are also currently evaluating other approaches to improve the results concerning normalization, dimension reduction, clustering, statistical analysis and data visualization.

Besides sequencing on an Illumina machine, we also offer the possibility to sequence on MinION (Oxford Nanopore Technologies), which produces long reads from DNA or RNA samples. We particularly focused on direct RNA sequencing (SQK-RNA002) and DNA sequencing (SQK-LSK110) on R9.4.1 flow cell. So far, we produced up to 15 million reads with a N50 ranging from 1 300 to 36 000 depending on the project. The basecalling is done by GPU version of Guppy in SUP model on two Nvidia Tesla V100 GPU cards. We have also developed a procedure for the de novo assembly of long-read only or hybrid long-read/short-read bacterial genomes.

A COLLABORATIVE methodology for MULTI-OMIC analysis

Florent Dumont[1], Luciana Oliveira[2], Claudine Deloménie[1], Emy Ponsardin[1], Firmin Akoumia[3], Guillaume Bernadat[1], Sylvia Cohen Kaminsky[3], Valérie Domergue[1]

[1] UMS IPSIT, Paris-Saclay University, 5 rue JB Clément 92196 Châtenay-Malabry

[2] ADLIN Science, 4 rue Pierre Fontaine, 91000, Evry-Courcouronnes, FR

[3] INSERM UMR-S 999, Paris-Saclay University, Pulmonary Hypertension. Hôpital Marie Lannelongue,133 avenue de la Résistance, 92350 Le Plessis-Robinson

Corresponding Author: florent.dumont@universite-paris-saclay Luciana@adlin-science.com

Reference paper :***https://www.pluginlabs-universiteparissaclay.fr/fr/entity/32d0d487-217f-4f59-b31d-0735653ec294/ipsit-ingenierie-et-plateformes-au-service-de-linnovation-therapeutique***
https://main.adlin-science.io/Absract
http://www.u999.universite-paris-saclay.fr/fr/

Abstract *Exploiting omic data has become a usual task within many medical and biological research laboratories. Indeed the increasing number of technological platforms and their attached services, the lessening of costs, and the publication of raw data impulse the creation of new standards in analysis methodology.Our analysis methodology relies on two axes: we carry out the bioinformatics analysis of data locally, comprehensively, and in one go to create a new base of analyzed data that is then biologically exploited and interpreted. Briefly, the workflow needs normalized data, statistical experimental design and symbol annotations. First, we apply quality control and unsupervised analysis on global data for all input factors. Then we apply analysis of variance model and compute post hoc pairwise comparisons tests with biological sense and their combination pattern with fold-change. The workflow then filters the results using a threshold gradient to make downstream analysis. All data subsets are used to generate venn and cluster analysis. All created lists of significant features from all subset analysis are then used for global enrichment analysis for functions, pathway, regulators, etc… using MSigDB[1] .*
We exploit the analyzed data thanks to a collaborative tool allowing putting together the know-how of bioinformaticians, bio-analysts, and biologists all involved in the project to optimize both interpretation and follow-up of the results. To enforce such methodology, we have developed the R package ***MOAL*** *(Multi-Omic Analysis at Lab), containing a simple function to make an omic analysis that will generate the basis of analyzed data[2]. To maintain a collaborative and transparent spirit in exploiting the results, we have created a fruitful private-public partnership with* ***ADLIN Science****[3]. This digital health-tech company is developing an innovative solution for data management and multi-omics analysis. ADLIN platform promotes the open science initiative, following fair principles to guarantee the security and traceability of scientific research. ADLIN workspace is a user-friendly, integrated environment that assists and guides the user in building and managing projects with different levels of complexity. Individual modules facilitate a wide range of tasks, such as data structuration, bioinformatics analysis, etc., carried out in parallel.*
We aim to present our integrated approach and its use through applied to a biological demonstrator generated in partnership with INSERM UMS-R 999. Here we explored the ***Implication of the NMDA*** *(N-methyl-D-aspartate)* ***receptor in the transcriptomic and proteomic PDGF*** *(Platelet derived growth factor subunit B)* ***Response of Pulmonary Vascular Smooth Muscle Cells from Patients with Pulmonary Arterial Hypertension****. This multi-omics analysis showed PDGF induced pro-proliferative gene and protein expression in pulmonary vascular smooth muscle cells that tended to normalize using NMDAR antagonists.*

References

[1] Arthur Liberzon & Co Molecular Signatures Data Base Bioinformatics 27(12):1739–1740 2011
[2] https://www.bioconductor.org/packages/release/bioc/html/moal.html (in progress)
[3] ADLIN Science. https://adlin-science.com/. Accessed: 2022-05-12.

ABiMS: Analysis and Bioinformatics for Marine Science

Loraine BRILLET-GUÉGUEN[1,2], Gildas LE CORGUILLÉ[1], Mark HOEBEKE[1], the ABiMS team, and Erwan CORRE[1]

[1] CNRS - Sorbonne Université - Plateforme ABIMS - Station Biologique de Roscoff, Place Georges Teissier, 29680, Roscoff, France
[2] Sorbonne Université, CNRS, Integrative Biology of Marine Models (LBI2M), Station Biologique de Roscoff (SBR), 29680, Roscoff, France

Corresponding Author: erwan.corre@sb-roscoff.fr

Des domaines tels que la biologie des systèmes, la modélisation des réseaux ou l'analyse des données NGS constituent un véritable défi en termes de calcul scientifique. Dans un contexte de production de données de biologie marine à haut débit et de traçabilité des analyses, le développement d'une infrastructure de calcul scientifique est une étape essentielle pour la production de connaissances.

La plateforme ABiMS (Analysis and Bioinformatics for Marine Science) de la Station biologique de Roscoff répond aux besoins des chercheurs et des chercheuses en biologie marine et, plus largement, des sciences de la vie. Créée en 2008, elle est l'une des plateformes nationales de l'Institut français de bioinformatique (IFB). Elle est également un Centre de Données et de Services *in situ* du pôle ODATIS, de l'infrastructure de recherche Data Terra. ABiMS fait partie du réseau IBISA et est membre du GIS BioGenouest.

La plateforme met au service de la communauté une infrastructure de calcul et de stockage (2500 CPU , 2.5 Po), ainsi qu'une palette de compétences, un catalogue de services organisés autour de 5 activités :

- Ingénierie logicielle : développement d'interfaces web couplées à des bases de données, centrées aussi bien sur des données de type séquence que sur des données d'observation
- Gestion de données : FAIRisation ou accompagnement à la FAIRisation, mise en accès de jeux de données FAIRisées, publication de jeux de données DOIsés dans des entrepôts thématiques
- Analyse bioinformatique : analyse de données (Assemblage et annotation de genome, transcriptomes, metagenomes etc, analyse de diversité) , etc.
- E-infrastructure : cluster de calcul, interfaces Galaxy, JBrowse, Apollo, R, espace de stockage ou outils bioinformatiques. Environnement pour l'analyse, l'annotation et l'hébergement de données de génomiques
- Support : demande de support pour l'utilisation des ressources de la plateforme (logiciels, données, etc.)
- Formation : formation aux méthodes et logiciels bioinformatiques

Pour assurer la qualité de nos services, nous avons mis en place un système de gestion de la qualité basé sur la norme ISO 9001, qui a été initialement certifié en 2014 et qui est approuvé à la norme ISO 9001:2015 depuis décembre 2017.
Grâce à ses nombreuses interactions avec les unités de recherche et en tant que membre du Réseau national des ressources informatiques de l'IFB et en tant que centre de données et de service , ABiMS est impliquée dans de nombreux projets de recherche (une vingtaine par an), dont les impacts nationaux et européens concernent les activités de bioanalyse, de développement logiciel et d'e-infrastructures.

The Migale bioinformatics core facility

Valentin LOUX[1,2], Mouhamadou BA[1,2], Hélène CHIAPELLO[1,2], Christelle HENNEQUET-ANTIER[1,2], Mahendra MARIADASSOU[1,2], Véronique MARTIN[1,2], Cédric MIDOUX[1,2,3], Olivier RUÉ[1,2], Valérie VIDAL[1,2] and Sophie SCHBATH[1,2]

[1] Université Paris-Saclay, INRAE, MaIAGE, Domaine de Vilvert, 78350, Jouy-en-Josas, France.

[2] Université Paris-Saclay, INRAE, BioinfOmics, MIGALE bioinformatics facility, Domaine de Vilvert, 78350, Jouy-en-Josas, France.

[3] Université Paris-Saclay, INRAE, PRocédés biOtechnologiques au Service de l'Environnement, 1 rue Pierre-Gilles de Gennes, CS10030, 92761, Antony, France.

Corresponding Author: valentin.loux@inrae.fr

https://migale.inrae.fr/

The Migale bioinformatics facility is a team of INRAE's MaIAGE research unit (Applied Mathematics and Computer Science, from Genome to the Environment). It has been providing services to the life sciences community since 2003.

Migale is an open platform, that offers four types of services ;

- an open infrastructure dedicated to life sciences data processing,
- dissemination of expertise in bioinformatics,
- design and development of bioinformatics applications,
- genomic, metagenomic and metatranscriptomic analysis.

Migale is part of the French Institute of Bioinformatics (IFB) and France Génomique projects. It has an ISO9001 certification and is also one of the four platforms which compose BioinfOmics, the national Research Infrastructure in bioinformatics of INRAE.

The poster will illustrate the platform services with examples chosen from recent achievements.

A complete description of Migale facility's service offer is available on its website : https://migale.inrae.fr

scAN10 : A reproducible and standardized pipeline for processing 10X single cell RNAseq data

Maxime Lepetit[1], Mirala Diana Ilie[2], Philippe Bertolino[2], Gerald Raverot[2], Olivier Gandrillon[1] and Franck Picard[1]
[1] Laboratoire de Biologie et Modélisation de la Cellule, ENSL, 46 allée d'Italie, 69364 LYON, FRANCE
[2] Centre de Recherche en Cancerologie de Lyon, 28 Rue Laennec, 69008 Lyon, FRANCE

Corresponding author: maxime.lepetit@ens-lyon.fr

Emergence of single cell transcriptomics has led to the challenge of developing data analysis pipelines that are both fully reproductible and modular while allowing interoperability across multiple systems and institutions, overcoming differences in hardware, operating systems and software versions.
We approached this challenge by designing scAN10 (single cell ANalysis of 10X data), a processing pipeline of 10X single cell RNAseq data, that is written in Nextflow and inherits the ability to be executed on most computational infrastructures.
This pipeline takes as input fastq files that are generated using the 10X Genomics protocol. In order to produce the filtered normalized count matrices, the fastq files are processed and trimmed by using Fastp [1], an ultrafast fastq processor that supports multi-threading with the aim of removing low quality bases. To compute feature barcode count matrix from trimmed fastq files, the pipeline gives the possibility to choose between two of the most popular single cell mappers: Kallisto-bustools [2] that perform an alignement free approach, so called pseudo-alignement or Cellranger [3] that performs a classical alignement approach with the underlying use of STAR. This step requires the input of genomic files (genome sequence and GTF annotation files) corresponding to the species of interest. In the case of human analysis these files can be downloaded automatically by specifying a version number available on the Ensembl database. To remove low quality cells, the count matrix is filtered by using the Seurat (R) package based on different criteria that the user can modify at the start of the pipeline. To normalize filtered data counts and ensure that any heterogenity observed is driven by biology and not by technical biases, the user is then allowed to use a Bayesian inference of gene expression states by filtering out the Poisson noise called Sanity [4]. Finally, a KNN based clustering with a Louvain modularity optimisation and the visualization of the data in a reduced dimension space with UMAP or T-SNE is done with the Seurat package. Alternatively this last step can be skipped allowing the user to use their own clustering method. Similarly, to avoid the introduction of layers of complexity and simplify the pipeline usage, the automatic annotation of clusters was not introduced. Users can annotate their dataset manually. In our case, we used the expression of sets of markers specific to our biological model.

We will illustrate the application of scAN10 to a clinical dataset of human pituitary gonadotroph tumours acquired from 6 patients (3 Men / 3 Women). This data were analysed using scAN10, with the perspective to better understand the cell heterogeneity and microenvironment composition of gonadotroph tumours, a frequent form (35% of total) of intracranial pituitary neoplasm with poor diagnostic, predictive and therapeutic clinical perspectives.
The git repository of scAN10 is avalaible at https://gitbio.ens-lyon.fr/LBMC/sbdm/scan10

We thank the PLASCAN institute for financing this project, and Laurent Modolo from the LBMC for his help for using the Nextflow language.

References

[1] Shifu Chen, Yanqing Zhou, Yaru Chen, and Jia Gu. fastp: an ultra-fast all-in-one fastq preprocessor. *Bioinformatics*, 34:i884–i890, 2018.

[2] Melsted P., Booeshaghi A.S., and Liu L. et al. Modular, efficient and constant-memory single-cell rna-seq preprocessing. *Nat Biotechnol*, (39):813–818, 2021.

[3] Zheng G., Terry J., and Belgrader P. et al. Massively parallel digital transcriptional profiling of single cells. *Nat Biotechnol*, (8):14049, 2017.

[4] Breda J., Zavolan M., and van Nimwegen E. Bayesian inference of gene expression states from single-cell rna-seq data. *Nat Biotechnol*, (39):1008–1016, 2021.

Single-cell Initiative of Institut Curie: presentation of technologies & bioinformatics resources

Louisa HADJ ABED[1,2,3], Rémi MONTAGNE[1,2,3], Pacôme Prompsy[4,5], Leanne De Koning[3,5], Céline Vallot[3,4,5,*], Nicolas Servant[1,2,3,*]

1 Institut Curie, PSL University, INSERM U900, 75005 Paris, France

2 Centre de Bio-Informatique (CBIO), MINES ParisTech, Institut Curie, PSL University, 75006 Paris, France

3 Institut Curie, PSL University, Single Cell Initiative, 75005 Paris, France

4 Institut Curie, PSL University, Sorbonne Université, CNRS UMR3244, Dynamics of Genetic Information, 75005 Paris, France

5 Institut Curie, PSL University, Translational Research Department, 75005 Paris, France

* Equally contributed

Corresponding Author: leanne.de-koning@curie.fr, nicolas.servant@curie.fr

The heterogeneity of healthy, tumor and immune cells plays a key role in normal and pathological development, as well as in response to treatment. Conventional technologies (in bulk) study a global population of cells and do not allow the detection of subpopulations of rare cells. To be able to distinguish the identity of each cell individually, high-throughput technologies at the scale of the single cell, have been under intense development for several years.

At the research center of Institut Curie, four core facilities from the Curie CoreTech network (Custom Single Cell Omics, Next Generation Sequencing (NGS ICGEX), Cytometry and Bioinformatics platforms) have decided to join forces and create the Single-Cell Initiative. This program aims to coordinate and share resources to optimize services and developments of single-cell technologies for the scientific community. The Single-Cell Initiative (i) benchmarks and adapts state-of-the art technologies, (ii) develops custom single-cell approaches, when necessary [1], and (iii) provides support for research teams with their single-cell experiments. This support includes cell characterization, cell sorting, single-cell isolation, sequencing and data analyses covering a large range of single-cell approaches from 3' and full-length RNA-seq, to single-cell epigenomics and spatial omics. Custom projects requiring dedicated microfluidic developments are also encouraged through the custom single-cell omics platform.

In addition to the experimental support, the single-cell initiative is developing bioinformatics pipelines for all supported single-cell technologies. The goal of these pipelines is to propose quality controls of the experiments as well as the pre-analysis steps such as data cleaning, filtering and first level analysis (e.g mapping, counting). These pipelines are based on the *Nextflow* framework, and developed following the FAIR principles of scalability, portability and re-usability. They can easily be used for both production environment (with the *geniac* framework [2]) and user applications, and support containers usage such as *conda*, *docker* and *singularity*.

We will present the Institut Curie Single-Cell initiative - its organization and the different supported technologies.

References

[1] Grosselin, K., Durand, A., Marsolier, J. *et al.* High-throughput single-cell ChIP-seq identifies heterogeneity of chromatin states in breast cancer. *Nat Genet* **51,** 1060–1066 (2019). https://doi.org/10.1038/s41588-019-0424-9

[2] Allain F., Roméjon J. La Rosa P. *et al.* Geniac: Automatic Configuration GENerator and Installer for nextflow pipelines. *Open Research Europe* 2021, 1:76 https://doi.org/10.12688/openreseurope.13861.1

[3] Prompsy, P., Kirchmeier, P., Marsolier, J. *et al.* Interactive analysis of single-cell epigenomic landscapes with ChromSCape. *Nat Commun* 11, 5702 (2020). https://doi.org/10.1038/s414667-020-19542-x

EDAM, life sciences ontology for data analysis and management.

Lucie Lamothe[1], Alban Gaignard[2], Mads Kierkegaard[3], Hager Eldakroury[4], Melissa Black[5], Bryan Brancotte[6], Jon Ison[1], Veit Schwämmle[3], Matúš Kalaš[7] , Hervé Ménager[6,1]

[1] CNRS, Institut Français de Bioinformatique, IFB-core, UAR 3601, Evry, France
[2] Nantes Université, CNRS, INSERM, l'institut du thorax, F-44000 Nantes, France
[3] Department of Biochemistry and Molecular Biology, University of Southern Denmark, Odense 5230, Denmark
[4] Outreachy intern (EDAM), Cairo (at the time of contribution)
[5] Outreachy intern (EDAM), São Paulo (at the time of contribution)
[6] Institut Pasteur, Université Paris Cité, Bioinformatics and Biostatistics Hub, F-75015 Paris, France
[7] University of Bergen, 5008 Bergen

Corresponding Author: lucie.lamothe@france-bioinformatique.fr

EDAM [1] is a domain ontology of data analysis and data management in life sciences. It comprises concepts related to analysis, modeling, optimization, and data life-cycle, and is divided into 4 main sections: topics, operations, data, and formats.

EDAM is used in numerous resources, for example bio.tools, Galaxy, CWL, Debian, BioSimulators, FAIRsharing, or the ELIXIR Europe training portal TeSS. Thanks to the annotations with EDAM, tools, workflows, standards, data, and learning materials are easier to find, compare, choose, and integrate. EDAM contributes to *open science* by allowing semantic annotation of research products, thus making them more understandable, findable, and comparable.

EDAM is continuously evolving and expanding by improving the implementation of links to external resources (including other ontologies), definitions, and the overall quality, or the addition of new concepts. EDAM is developed in a participatory and transparent fashion, within a broad and growing community of contributors. This development model, based on the contribution of a large number of scientific experts, therefore comes with its own set of challenges.

To ease the contribution processes, users can explore graphically the ontology and its most useful features using the EDAM Browser [2] web interface.

To streamline and accelerate the evolution of EDAM, we have developed and integrated a set of tools that automate the quality control and release process for the ontology. In addition to ensuring the global consistency of EDAM, it enforces edition best practices both at the syntactic and semantic levels. These tools have been integrated in a *continuous integration* (CI) pipeline, automated using GitHub Actions in the source-code repository.

These tools participate in the improvement of EDAM's contribution process and visibility by the community.

References

[1] Ison, J., Kalaš, M., Jonassen, I., Bolser, D., Uludag, M., McWilliam, H., Malone, J., Lopez, R., Pettifer, S. and Rice, P. (**2013**). EDAM: an ontology of bioinformatics operations, types of data and identifiers, topics and formats. *Bioinformatics*, **29**(10): 1325-1332. DOI: 10.1093/bioinformatics/btt113 *Open access*

[2] Brancotte, B., Blanchet, C. and Ménager, H. (**2018**). A reusable tree-based web-visualization to browse EDAM ontology, and contribute to it. *J. Open Source Softw.*, **3**(27): 698. DOI: 10.21105/joss.00698 *Open access*

REPET evolutions: faster and easier

Mariène WAN [1,2], Johann CONFAIS [1,2] and Hadi QUESNEVILLE [1,2]

[1] Université Paris-Saclay, INRAE, URGI, 78026, Versailles, France

[2] Université Paris-Saclay, INRAE, BioinfOmics, Plant bioinformatics facility, 78026, Versailles, France

Corresponding Author: mariene.wan@inrae.fr

Transposable elements (TEs) are major players of structure and evolution of eukaryote genomes. Thanks to their ability to move around and to replicate within genomes, they are probably the most important contributors to genome plasticity. Their detection and annotation are considered essential and must be undertaken in any genome sequencing project.

The REPET package [1, 2] integrates bioinformatics pipelines dedicated to detect, annotate and analyze TEs in genomic sequences. The two main pipelines are (i) TEdenovo, that search for interspersed repeats, build consensus sequences and classify them [3] according to TE features and (ii) TEannot, which mines a genome with a library of TE sequences, for instance the one produced by the TEdenovo pipeline, to provide TE annotations.

The REPET package is in continuous improvement for speed by parallelizing several key bottleneck steps. In addition, several strategies which reduce the time required for analyzing large genome have been tested. With the speed improvement and adapted strategies, REPET is now able to annotate and analyze genomes such as the maize with more than 85% of TEs on a 2.3 Gb genome [4] on current computer cluster.

With this tool, the PlantBioinfoPF platform ensures a TE annotation service. Indeed, we are now able to propose an automatic TE annotation of good quality through a process called "Repet-Factory". This process uses the REPET software suite with parameters optimized for TE detection specificity and computing time. This process is capable of successively annotate several genomes in batches with the required traceability and reproducibility of the analyzes.

Moreover, a Virtual Research Environment (VRE) for TE annotation and its analysis has been developed on Virtual Machines (VM). An ansible script instantiate VMs with all packages and tools required for a complete genome annotation with the REPET package. This script allows this VRE to be easily re-instantiated in other infrastructures which greatly simplify the REPET package installation with all its required dependencies. We also simplified the distribution of REPET to increase its availability and portability to users, by developing a Docker image of REPET (https://hub.docker.com/r/urgi/docker_vre_aio).

The REPET tool is a cornerstone of the platform. In addition to its use in the genome TE annotation service and its availability for download, it is also the basis of the RepetDB [5] database (https://urgi.versailles.inrae.fr/repetdb) hosted by the platform which provides libraries of reference TE sequences for more than 50 species.

References

1. Flutre T, Duprat E, Feuillet C, Quesneville H (2011) Considering Transposable Element Diversification in De Novo Annotation Approaches. PLoS ONE 6(1): e16526. https://doi.org/10.1371/journal.pone.0016526

2. Quesneville H, Bergman CM, Andrieu O, Autard D, Nouaud D, Ashburner M, et al. (2005) Combined Evidence Annotation of Transposable Elements in Genome Sequences. PLoS Comput Biol 1(2): e22. https://doi.org/10.1371/journal.pcbi.0010022

3. Hoede C, Arnoux S, Moisset M, Chaumier T, Inizan O, Jamilloux V, et al. (2014) *PASTEC: An Automatic Transposable Element Classification Tool.* PLoS ONE 9(5): e91929. https://doi.org/10.1371/journal.pone.0091929

4. V. Jamilloux, J. Daron, F. Choulet and H. Quesneville, "*De Novo Annotation of Transposable Elements: Tackling the Fat Genome Issue*," in Proceedings of the IEEE, vol. 105, no. 3, pp. 474-481, March 2017, doi: 10.1109/JPROC.2016.2590833.

5. Amselem, J., Cornut, G., Choisne, N., Alaux, M., Alfama-Depauw, F., Jamilloux, V., Maumus, F., Letellier, T., Luyten, I., Pommier, C., Adam-Blondon, A. F., & Quesneville, H. (2019). RepetDB: a unified resource for transposable element references. Mobile DNA, 10, 6. https://doi.org/10.1186/s13100-019-0150-y

Comparison of Stacks and a custom pipeline for RADseq analysis

Enora Geslain[1], Álvaro Cortés Calabuig[2], Sarah M. Maes[1], Gregory E. Maes[2] and Filip A.M. Volckaert[1]

[1] Laboratory of Biodiversity and Evolutionary Genomics, Charles Deberiotstraat 32, PO Box 2439, B-3000, Leuven, Belgium

[2] Genomics Core Leuven, Herestraat 49, PO Box 606, B-3000, Leuven, Belgium

Corresponding Author: enora.geslain@kuleuven.be

1. Introduction

Restriction-site associated DNA sequencing (RADSeq) is a technique to scan complete genomes of organisms without sequencing them entirely. RADSeq is commonly used to identify thousands of single nucleotide polymorphisms (SNPs) randomly distributed across the genome for applications in population genomics. Different tools exist to search for SNPs from RADSeq data. Here, we compare two pipelines: a custom pipeline (using among others the GBSx [1] tool for demultiplexing) and the Stacks pipeline developed by Rochette and Catchen [2].

2. Approach

Both pipelines were applied on a dataset of 192 individual fish (polar cod; *Boreogadus saida*) from 10 different sampling regions in the Arctic Ocean. The samples were paired-end sequenced on a single run of an Illumina NovaSeq6000 device. The custom pipeline consists of the following steps: demultiplexing using GBSx v. 1.3, merging of paired-end reads with Flash v. 1.2.11, read mapping using Bowtie2 v. 2.3.4.3 or BWA v. 0.7.17, annotating read groups with Picard v. 2.18.23 and finally SNP calling using Freebayes v. 1.3.2. Stacks pipeline v. 2.5 was applied following the recommendations of Rochette and Catchen. To compare the results of both pipelines, VCFtools v. 0.1.16 was used to filter the VCF files because the populations module of Stacks is not applicable to VCF files generated with other tools. Finally, we performed a Principal Component Analyses (PCA) for the different VCF files using R (adegenet v. 2.1.5) to have a better insight in the differences of both pipelines.

3. Results

GBSx retrieved an average of 0.34% (= 1,496,545) more paired reads compared to the process_radtags command of Stacks. We also observed a difference between both mappers, with BWA having a higher alignment rate than Bowtie2 in both pipelines. A difference between both pipelines is also present as more reads are aligned in Stacks than in the custom pipeline.

Removing indels had an effect only with the custom pipeline because, unlike gstacks, Freebayes retrieved all variants, not only SNPs. We filtered the SNPs using a maximum of 75% missing data per individual, a minimum allele frequency (MAF) of 0.05, a minimum allele count (MAC) of 5 and a maximum of 20% missing data per site. We retrieved with the Stacks pipeline more SNPs when using Bowtie2, with only 22% SNPs in common with the BWA dataset.

Finally, we detected with the PCAs a possible lane effect in the Bowtie2 datasets. Otherwise, the results are comparable for all Stacks datasets: one population clustering apart from the others, which was expected given its remote location (the Freebayes results will be integrated in the final analysis of the poster).

4. Discussion

Choosing tools correctly when doing RADSeq analysis is essential because each tool can have an impact on downstream analysis. Here, the best tool seems to be Bowtie2 for the mapping. BWA was developed for human genome analysis, which may explain its reduced efficiency for other organisms. However, further filtering might be required to remove the lane effect detected with Bowtie2.

References

1. K. Herten. GBSX: a toolkit for experimental design and demultiplexing genotyping by sequencing experiments. *BMC Bioinformatics* 16:73, 2015.

2. N. Rochette. Stacks 2: Analytical methods for paired‐end sequencing improve RADseq‐based population genomics. *Molecular Ecology* 28(21):4737-4754, 2019.

ePeak: from replicated chromatin profiling data to epigenomic dynamics

Maëlle Daunesse[1,2,*], Rachel Legendre[1,*], Hugo Varet[1], Adrien Pain[1] and Claudia Chica[1]
[1] Bioinformatics and Biostatistics Hub, Institut Pasteur, Université de Paris Cité, F-75015, Paris, France
[2] Flicek Research Group, European Bioinformatics Institute, CB10 1SD, Cambridge, UK

Corresponding author: claudia.chica@pasteur.fr

1 Introduction

We present ePeak[1], a Snakemake-based pipeline for the identification and quantification of reproducible peaks from raw ChIP-seq, CUT&RUN and CUT&Tag epigenomic profiling techniques. It also includes a statistical module to perform tailored differential marking and binding analysis with state of the art methods. ePeak streamlines critical steps like the quality assessment of the immunoprecipitation, spike-in calibration and the selection of reproducible peaks between replicates for both narrow and broad peaks. It generates complete reports for data quality control assessment and optimal interpretation of the results. We advocate for a differential analysis that accounts for the biological dynamics of each chromatin factor. Thus, ePeak provides linear and nonlinear methods for normalisation as well as conservative and stringent models for variance estimation and significance testing of the observed marking/binding differences. Using a published ChIP-seq dataset, we show that distinct populations of differentially marked/bound peaks can be identified. We study their dynamics in terms of read coverage and summit position, as well as the expression of the neighbouring genes. We propose that ePeak can be used to measure the richness of the epigenomic landscape underlying a biological process by identifying diverse regulatory regimes.

The pipeline is freely available at https://gitlab.pasteur.fr/hub/ePeak/ under GNU General Public Licence.

References

[1] Varet H Pain A Chica C Daunesse M, Legendre R. epeak: from replicated chromatin profiling data to epigenomic dynamics. *NAR Genomics and Bioinformatics*, page 10.1093/nargab/lqac041, 2022.

*These authors contributed equally to this work

CEA JupyterHub platform for multi-omics data analysis

Solène Mauger[1], Florian Jeanneret[1], Pauline Bazelle[1], Christophe Battail[1] and KATY consortium[2]

[1] Laboratoire Biologie et Biotechnologies pour la Santé, IRIG, UMR 1292 INSERM-CEA-UGA, Univ. Grenoble Alpes, 38000 Grenoble, France.

[2] https://katy-project.eu, Européen Unions's Horizon 2020 research ans innovation programme, Grant agreement No 101017453

Corresponding Author: christophe.battail@cea.fr

CEA Grenoble has recently set up a JupyterHub infrastructure (https://jupyter.org/hub) capable of putting the power of Notebooks (1) in the hands of researchers and students. This shared work environment allows Data Scientists and Biologists to perform their data exploration on shared computing resources. This paradigm shift allows the implementation of good practices associated with Open Science (https://www.fosteropenscience.eu/) which are the reproducibility of analyses and the FAIR principles (2).

The Notebooks are developed by bioinformaticians and are deposited on the CEA JupyterHub platform in order to allow easy access to the various members, researchers, engineers and students of biology research teams. This new data analysis environment is currently composed of Notebooks allowing to carry out differential analyses of gene expression, ontological enrichment studies, using different statistical methods and several reference databases, and analyses to identify master transcriptional regulators. In order to facilitate access to this work environment by non-bioinformaticians, the user starts its analysis by a Notebook with a graphical interface, implemented with the Python Dash software library, allowing to enter input data and to select the analyses to perform associated with the appropriate parameters, without having to modify the Notebook code.

In compliance with the FAIR approach, the Jupyter Notebooks developed within the framework of this CEA JupyterHub platform will be accessible from a public Gitlab repository.

References

[1] Kluyver T et al. and Jupyter development team. Jupyter Notebooks – a publishing format for reproducible computational workflows. In *Positioning and Power in Academic Publishing: Players, Agents and Agendas.* IOS Press. pp. 87-90. 2016. doi: 10.3233/978-1-61499-649-1-87.

[2] Wilkinson MD et al. The FAIR Guiding Principles for scientific data management and stewardship. Sci Data. 2016 Mar 15;3:160018. doi: 10.1038/sdata.2016.18.

Creation of an integrated molecular dynamics workflow on the Galaxy platform : Characterization of aquaporin pores

Agnès-Elisabeth Petit[1], Jean-Stéphane Venisse[2], Philippe Label[2] and Nadia Goué[1]

[1] AuBi platform, Clermont Auvergne Mesocenter, UCA , 7 Avenue Blaise Pascal, 63170 AUBIERE, France

[2] UMR PIAF, Université Clermont Auvergne, INRAE, PIAF, 63000 Clermont-Ferrand, France

Corresponding Author: nadia.goue@uca.fr

Galaxy is an international bioinformatics platform for biologists [1] . So far, the Galaxy team has adapted molecular dynamics tools which are mainly tools to create the prerequisites of a simulation or to run a simulation. In our case, this simulation step was done but the tools to finalize our analysis were missing. This is why tools have been developed and integrated in Galaxy. This integration of a succession of internal tools in the form of a Galaxy workflow is intended to help biologists and would benefit from the high performance computing facilities connected to the Galaxy webservice.

Tools developed here aim at studying the structure the structure of 102 aquaporin trajectories using a molecular dynamics approach. This approach requires to take into account the molecular scale (Ångströms) of the proteins and the time step (nanosecond). In total, we speak of a simulated trajectory of 100 ns to model the transport of a water molecule [2]. In order to optimize the computational time on a trajectory, each trajectory is divided into several sub-trajectories and the pore diameter calculations are performed for each sub-trajectory. The resulting data are then compiled in a table before being visualized in graphical form.

This workflow is designed to work on aquaporin trajectories. Aquaporins are transmembrane proteins that transport water. In addition, an aquaporin is a tetramer composed of four protomers. Each protomer has six transmembrane alpha helices connected by extramembrane loops that structure into a central pore. In addition, each protomer is hourglass-shaped and has two sites consisting of 3 successive amino acids, Asparagine - Proline - Alanine (NPA) and an aromatic arginine site (arR) [3]. The NPA sites form an electrostatic barrier preventing excess protons from entering the cell. The arR site is composed of 4 amino acids that form a constriction inside the pore of each protomer. This constriction prevents large particles from passing and also regulates the amount of water that can pass through the transmembrane space at any given time. Our worlflow allows us to calculate the pore diameter at this constriction. Recent advances in pore diameter characterization of aquaporin complexes, from manipulation of molecular modeling files to visualization of results, will be presented here.

References

[1] V. Jalili *et al.*, « The Galaxy platform for accessible, reproducible and collaborative biomedical analyses: 2020 update », *Nucleic Acids Research*, vol. 48, nº W1, p. W395-W402, juill. 2020, doi: 10.1093/nar/gkaa434.

[2] R. O. Dror, R. M. Dirks, J. P. Grossman, H. Xu, et D. E. Shaw, « Biomolecular Simulation: A Computational Microscope for Molecular Biology », *Annu. Rev. Biophys.*, vol. 41, nº 1, p. 429-452, juin 2012, doi: 10.1146/annurev-biophys-042910-155245.

[3] J.-S. Venisse *et al.*, « Genome-Wide Identification, Structure Characterization, and Expression Pattern Profiling of the Aquaporin Gene Family in Betula pendula », *IJMS*, vol. 22, nº 14, p. 7269, juill. 2021, doi: 10.3390/ijms22147269.

Supports for imaging projects toward Open Science at AuBi platform

Mateo HIRIART[1], David GRIMBICHLER[1] , Laurent BOURI[2], Pierre POUCHIN[3], Sophie DESSET[3], Julien SEILER[2], Pierre PEYRET[4], Antoine MAHUL[1], Valérie LEGUÉ[5] and Nadia GOUÉ[1]

[1] AuBi Platform, Clermont Auvergne Mesocenter, UCA, 7 avenue Blaise Pascal, 63 178 Aubière, France

[2] IGBMC, 1 Rue Laurent Fries, 67400, Illkirch-Graffenstaden, France

[3] iGReD, UMR UCA CNRS 6293 Inserm U1103, 28 place Henri Dunant, 63 001 Clermont-Ferrand, France

[4] MEDIS, UMR 454 UCA INRAE, 28 Place Henri Dunant,63 001 Clermont-Ferrand, France

[5] PIAF, UMR 547 UCA INRAE, 1 Impasse Amélie Murat ,63 178 Aubière, France

Corresponding Author: nadia.goue@uca.fr

1. Introduction

Hosted by the Clermont Auvergne Mesocenter [1], the Auvergne Bioinformatics (AuBi) platform is member of the French Institute of Bioinformatics (IFB) [2]. We aim at sharing expertise and knowledge in large-scale data analysis with computing and storage facilities for biology and health research laboratories located at Clermont Auvergne University (UCA). Among axis of interest is imaging for which OMERO [3] a tool for data visualization and metadata storage as well as a bio-analysis platform Galaxy [4] have been deployed.

2. A web portal to orchestrate the tools around the FAIR data

With the recent advances towards Open Science, data flows, tools and workflow analysis need to be orchestrated in order to help scientists to navigate across projects and collaborate together. To initiate this structuration around the data projects and based on FAIR principles, we started working on a web portal accessible from the Mesocenter hub. This portal shall give access to several services : (1) a unique corresponding address to help scientists feeding Data Management Plan (DMP) and HAL publication referencing system leaded by University library ; (2) mature web services for meta-data and image storing system, as well as images workflow analysis ; (3) a high performance calculation and storage infrastructure provided by the Mesocenter. Our objective is to be able to link these tools through data projects. This is why we are working together with OpenLink team [5] to develop a service meeting our biologists and bioinformatics communities, the Mesocenter and UCA requirements.

3. Conclusion

This Open Science project gives an opportunity for laboratories associated with AuBi platform to get involved in national IFB projects such as MuDiS4LS. This is the case for iGReD [6] and its confocal microscopy facility CLIC (CLermont Imagerie Confocal) by contributing to the Implementation Study for the elaboration of the structure DMP dedicated to imaging. This is challenging laboratories with imaging projects to mutualize storage resources like the storage server bought in 2021 all together with four Mixed Research Units : GDEC, iGReD, PIAF and UNH [6]. This project is initiated on imaging data and once on track, other data types will be orchestrated and linked together.

Keywords Open Science, Imaging, OpenLink, OMERO, GALAXY

References

[1] www.mesocentre.uca.fr
[2] www.france-bioinformatique.fr
[3] http://www.openmicroscopy.org/
[4] Afgan E. *et al*. The Galaxy platform for accessible, reproducible and collaborative biomedical analyses : 2018 update. 46(W1) : W537-W544, 2018.
[5] https://gitlab.com/igbmc/openlink
[6] https://ressources.france-bioinformatique.fr/fr/plateformes/aubi

Development of a pipeline integrating single-cell omic sequencing and phenotypic imaging analyses

Coline GARDOU[1], Gaël BLIVET-BAILLY[1] and Mathieu BAHIN[1]

[1] Minos Biosciences, ESPCI Paris PSL, 10 rue Vauquelin, 75005, Paris, France

Corresponding Author: coline.gardou@minosbiosciences.com

Minos Biosciences develops an instrument that offers the unique ability to directly combine sequencing-based multi-omic analysis and image-based phenotypic analysis at high-throughput single-cell resolution. This is enabled using a breakthrough microfluidics concept, as well as innovative approaches in imaging and molecular biology, to isolate, image and barcode cells. Together with the instrument, a bioinformatics solution will be provided to analyse these novel multimodal data.

It is necessary to develop a robust and efficient data analysis pipeline to analyze Minos Biosciences specific coupled omic and phenotypic datasets. This pipeline will be composed of published open source tools (like scanpy [1]) and new ones developed to take advantage of Minos Biosciences peculiarities. Firstly, the doublet or dying cells detection is assessed with certainty from the acquired images instead of having to be guessed from sequencing data. Secondly, the additional information obtained by immunofluorescence and features detection on these same cells imaging is used to refine the identification of cell populations.

The pipeline will integrate both omics and image analyses steps. The omics part consists of different steps: demultiplexing (by STARsolo [2]), quality control, batch effect correction [3], clustering and cell type annotation. The demultiplexing step will take advantage of imaging by only considering the relevant barcodes, the ones associated with actual isolated cells. The whole process outputs will be synthesized in an automated analysis report using MultiQC [4] and mainly lie in an augmented count matrix. The nf-core bioinformatics community [5] provides many standardized, portable and reproducible pipelines using Nextflow [6] as well as templates to create custom pipelines. Our pipeline, stemming from a nf-core template, is modular so that parts of it could be integrated to pipelines dealing with other omics data. To guarantee reproducibility, Singularity containers [7] are used on the fly to run the steps of the pipeline.

The omics part of the pipeline, for scRNA-seq, is complete from raw FASTQ files to cell population annotation while the phenotyping part is still under development. The integration of both modalities still needs to be studied. For a dataset of 67M reads, the pipeline execution time is 9m20s and reaches a maximum RAM usage of 30G RAM on a recent server with 12 CPU allocated for the demultiplexing step. In the coming months, tests on new, varied, datasets will help strengthen its robustness.

Minos Biosciences solution will provide highly accurate insights into complex cell populations and their dynamics.

References

[1] Wolf, F., Angerer, P. & Theis, F. SCANPY: large-scale single-cell gene expression data analysis. Genome Biol 19, 15 (2018). https://doi.org/10.1186/s13059-017-1382-0

[2] Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, Batut P, Chaisson M, Gingeras TR. STAR: ultrafast universal RNA-seq aligner Bioinformatics. 2013 Jan 1;29(1):15-21. doi: 10.1093/bioinformatics/bts635. Epub 2012 Oct 25. PubMed PMID: 23104886; PubMed Central PMCID: PMC3530905.

[3] Tran, H.T.N., Ang, K.S., Chevrier, M. et al. A benchmark of batch-effect correction methods for single-cell RNA sequencing data. Genome Biol 21, 12 (2020). https://doi.org/10.1186/s13059-019-1850-9

[4] Ewels P, Magnusson M, Lundin S, Käller M. MultiQC: summarize analysis results for multiple tools and samples in a single report. Bioinformatics. 2016 Oct 1;32(19):3047-8. doi: 10.1093/bioinformatics/btw354. Epub 2016 Jun 16. PubMed PMID: 27312411; PubMed Central PMCID: PMC5039924.

[5] Ewels PA, Peltzer A, Fillinger S, Patel H, Alneberg J, Wilm A, Garcia MU, Di Tommaso P, Nahnsen S. The nf-core framework for community-curated bioinformatics pipelines. Nat Biotechnol. 2020 Mar;38(3):276-278. doi: 10.1038/s41587-020-0439-x. PubMed PMID: 32055031.

[6] Di Tommaso P, Chatzou M, Floden EW, Barja PP, Palumbo E, Notredame C. Nextflow enables reproducible computational workflows. Nat Biotechnol. 2017 Apr 11;35(4):316-319. doi: 10.1038/nbt.3820. PubMed PMID: 28398311.

[7] Kurtzer GM, Sochat V, Bauer MW. Singularity: Scientific containers for mobility of compute. PLoS One. 2017 May 11;12(5):e0177459. doi: 10.1371/journal.pone.0177459. eCollection 2017. PubMed PMID: 28494014; PubMed Central PMCID: PMC5426675.

PitViper: a software for comparative meta-analysis and annotation of functional screening data

Paul-Arthur Meslin[1], Lois Kelly[1], Alexandre Puissant[1] and Camille Lobry[1]
Université Paris Cité, Génomes, Biologie Cellulaire et Thérapeutique, INSERM U944, CNRS UMR 7212, Paris 75110, France

Corresponding author: paul-arthur.meslin@inserm.fr and camille.lobry@inserm.fr

1 Abstract

Functional screening protocols, such as shRNA and CRISPR/Cas9, allow for detection of elements of genomes whose functionality is essential or incompatible with the maintenance of a phenotype of interest. Each functional screening protocol has been followed by the development of several bioinformatic algorithms dedicated to the identification of essential elements by building statistical models taking into account their own specificities. Here, we present, PitViper, an user-friendly bioinformatics software for integrative screening data analysis, that we developed with the aim of: (i) accept standard input formats for *de novo* and reanalysis of functional screening experiments; (ii) perform reads quantification, normalization and essentiality analysis using current gold standard tools and additional methods; (iii) allow for multiple methods integration; (iv) generate interactive and customable reports for data annotation and analysis, with publication-ready figures. PitViper is a comprehensive and convenient tool for functional screening data analysis for both bioinformaticians and researchers without extensive computer expertise.

PitViper was developed to analyze shRNA, CRISPR/Cas9 and CRISPR/dCas9 screens and supports most common file formats of screening experiments as starting input, such as raw unaligned sequences as FASTQ file, aligned sequences as BAM file or count matrix as a text file. PitViper workflow is organized as follows: (i) reads are mapped to a reference library of guides/hairpins, (ii) read abundance is quantified for each guide/hairpin in the library using mapping results, (iii) essential elements are identified using changes in read abundance between conditions, and then (iv) results are integrated before annotation and visualization. Mapping and quantification of reads is based on Bowtie2 and/or MAGeCKcount [1]. Essentiality analysis is performed using several reference methods such as MAGeCK MLE [2] and RRA [1], BAGEL [3], CRISPhieRmix [4] in addition of our own methods and work both for genes or genomic positions automatically annotated with proximal genes. PitViper consists of: first, a bioinformatics pipeline produced using Snakemake, a workflow management system for creating reproducible and scalable data analyses. Then, a command line interface written in Python3 was developed to facilitate the use in an automated and reproducible way. This command-line interface allow easy reanalysis of previously generated results. Finally, a graphical user-interface allows users to run PitViper with chosen parameters. All dependencies can be installed with the Conda package manager using a YAML file with specifications of all dependencies or with Docker. Actionable Jupyter Notebook reports are automatically generated for dynamic viewing and exporting of results as HTML pages. Interactivity is achieved using Altair, a statistical visualization library for Python, and the python library Ipywidgets for modules parameterization. Reports are customizable and contain quality control, single tool results or multiple method integration and data annotation with external tools/databases, such as EnrichR, Genamania or Depmap.

References

[1] Wei Li, Han Xu, Tengfei Xiao, Le Cong, Michael I Love, Feng Zhang, Rafael A Irizarry, Jun S. Liu, Myles Brown, and X. Shirley Liu. MAGeCK enables robust identification of essential genes from genome-scale CRISPR/Cas9 knockout screens. *Genome biology*, 15(12):554, 2014.

[2] Wei Li, Johannes Köster, Han Xu, Chen Hao Chen, Tengfei Xiao, Jun Shirley Liu, Myles Brown, and X. Shirley Liu. Quality control, modeling, and visualization of CRISPR screens with MAGeCK-VISPR. *Genome Biology*, 16(1), dec 2015.

[3] Traver Hart and Jason Moffat. BAGEL: A computational framework for identifying essential genes from pooled library screens. *BMC Bioinformatics*, 17(1), apr 2016.

[4] Timothy P. Daley, Zhixiang Lin, Xueqiu Lin, Yanxia Liu, Wing Hung Wong, and Lei S. Qi. CRISPhieRmix: A hierarchical mixture model for CRISPR pooled screens. *Genome Biology*, 19(1), oct 2018.

UseGalaxy.fr: a Galaxy server for the French bioinformatics community

Anthony Bretaudeau[1,2,3], Thomas Chaussepied[2,3], Lain Pavot[4], Julien Seiler[5,3] and Gildas Le Corguillé[6,3]

[1] IGEPP, INRAE, Institut Agro, Univ Rennes, 35000, Rennes, France
[2] Plate-forme GenOuest, Univ Rennes, Inria, CNRS, IRISA, F-35000 Rennes
[3] CNRS, Institut Français de Bioinformatique, IFB-core, UMS 3601, Évry, France
[4] INRAE, Unité Nutrition Humaine, PFEM, UMR 1019, 63122 Saint Genès Champanelle
[5] CNRS UMR 7104, BiGEst, IGBMC, 1 rue Laurent Fries, 67404 Illkirch Cedex
[6] Sorbonne Université, CNRS, FR2424, ABiMS, Station Biologique, 29680, Roscoff, France

Corresponding Author: anthony.bretaudeau@inrae.fr

Galaxy [1] is an open-source web-based computational portal, primarily specialized in bioinformatics. It enables accessible, reproducible, transparent and FAIR (Findable Accessible Interoperable Reusable) data analysis. Using it, scientists can perform data treatments, including single tool executions, more complex and data-intensive scientific workflows, or even custom code execution using on-demand interactive environments (RStudio, Jupyter, RShiny, ...). Galaxy also allows to manage scientific data by easily sharing and publishing scientific results and workflows, with customisable reporting and visualisations. This web portal makes computational biology accessible to scientists without requiring computer programming skills.

The French Institute of Bioinformatics (IFB) offers various services to facilitate access to life science data processing. Part of these services is based on its IFB Core Cluster, a Slurm based high-performance computing infrastructure, composed of 4300 cores, 20 TB of RAM and 2PB of storage. Built on top of the Core Cluster, UseGalaxy.fr, the French national Galaxy instance, is available for free. The target audience includes any scientist from the national biology and bioinformatics community. Officially launched in 2020, it is operated by IFB, and it already provides access to 1850 tools to more than 2300 registered users.

Following the international UseGalaxy.* [2] guidelines, administration of usegalaxy.fr is based on a IaC (Infrastructure as Code) model: every aspect of the portal is managed using Ansible (developed collaboratively with the international community), deposited on public GitLab repositories [3]. As such, in a collaborative manner, UseGalaxy.fr is continuously being improved to address the needs of the community.

While many local Galaxy instances exist in the French community, the intention of UseGalaxy.fr is to propose a national-scale Galaxy instance with robust computing and storage capacities, and federated human resources for the development and support activities. Members of multiple regional platforms (Rennes, Roscoff, Strasbourg, ...) are regular contributors to this project, and new members are always welcome.

In addition to the global UseGalaxy.fr portal, five thematic subdomains have been set up (Workflow4Metabolomics, Covid19, ProteoRE, Metabarcoding, Ecology); allowing users to find a more focused choice of specific tools. It also offers a complete training platform, by using the Galaxy Training Network (a collection of tutorials for users), and since recently the TIaaS (Training Infrastructure as a Service), a tool to assist the trainers: it allows to reserve dedicated computing resources on the IFB-core cluster (to reduce waiting time on the day of training), and to track progress of students during the event.

Finally, UseGalaxy.fr is supporting several IFB flagship health projects: EMERGEN (genomic surveillance of SARS-CoV-2), ABRomics (research and monitoring project on antibiotic resistance) through its API and workflow system.

1. Enis Afgan *et al.*, *The Galaxy platform for accessible, reproducible and collaborative biomedical analyses: 2018 update*, Nucleic Acids Research, Volume 46, Issue W1, 2 July 2018, Pages W537–W544, doi:10.1093/nar/gky379.
2. The Galaxy Community. *Galaxy Community Update. BCC2020*. https://vimeo.com/440168116
3. https://gitlab.com/ifb-elixirfr/usegalaxy-fr

SCHNAPPs - Single Cell sHiNy APPlication(s)

Bernd Jagla[12], Valentina Libri[1], Claudia Chica[2] Vincent Rouilly[3], Sébastien Mella[12], Michel Puceat[12] and Milena Hasan[1]

[1] Institut Pasteur, Université de Paris, Cytometry and Biomarkers UTechS, F-75015 Paris, France
[2] Institut Pasteur, Université de Paris, Bioinformatics and Biostatistics Hub, F-75015 Paris, France
[3] Datactix, 40 rue Neuve, 33000, Bordeaux, France
[4] Aix-Marseille University, INSERM U-1251, MMG, France

Corresponding author: bernd.jagla@pasteur.fr

Single-cell RNA-sequencing (scRNAseq) experiments are becoming a standard tool for bench-scientists to explore the cellular diversity present in all tissues. Data produced by scRNAseq is technically complex and requires analytical workflows that are an active field of bioinformatics research, whereas a wealth of biological background knowledge is needed to guide the investigation. Thus, there is an increasing need to develop applications geared towards bench-scientists to help them abstract the technical challenges of the analysis so that they can focus on the science at play. It is also expected that such applications should support closer collaboration between bioinformaticians and bench-scientists by providing reproducible science tools.

We present SCHNAPPs, a Graphical User Interface (GUI), designed to enable bench-scientists to autonomously explore and interpret scRNAseq data and associated annotations. The R/Shiny-based application allows following different steps of scRNAseq analysis workflows from Seurat or Scran packages: performing quality control on cells and genes, normalizing the expression matrix, integrating different samples, dimension reduction, clustering, and differential gene expression analysis. Visualization tools for exploring each step of the process include violin plots, 2D projections, Box-plots, alluvial plots, and histograms. An R-markdown report can be generated that tracks modifications and selected visualizations. The modular design of the tool allows it to easily integrate new visualizations and analyses by bioinformaticians. We illustrate the main features of the tool by applying it to the characterization of T cells in a scRNAseq and Cellular Indexing of Transcriptomes and Epitopes by Sequencing (CITE-Seq) experiment of two healthy individuals.

The IFB Core Cluster : an open HPC resource for all biologists and the breeding ground of the IFB National Network of Computational Resources (NNCR)

David Benaben[1,2,3], Anthony Bretaudeau[1,4,5], Philippe Bordon[1,6], Thomas Chaussepied[1,5], Nicole Charrière[1,5], François Gerbes[1,7], Jean-François Guillaume[1,8], Jean-Christophe Haessig[1,9], Didier Laborie[1,6], Guillaume Seith[1,9], Julien Seiler[1,9*] and Gildas Le Corguillé[1,7*]

[1] IFB/Institut Français de Bioinformatique, CNRS UMS 3601, Génoscope, 91057, ÉVRY, France
[2] UMR Biologie du Fruit et Pathologie, Université de Bordeaux, INRAE, F-33140 Villenave d'Ornon, France
[3] Bordeaux Metabolome, MetaboHUB, PHENOME-EMPHASIS, F-33140 Villenave d'Ornon, France
[4] IGEPP, INRAE, Institut Agro, Univ Rennes, 35000, Rennes, France
[5] Plate-forme GenOuest, Univ Rennes, Inria, CNRS, IRISA, F-35000 Rennes
[6] GenoToul-Bioinfo, INRAE, 24 chemin de Borde-Rouge, 31320, Auzeville-Tolosan, France
[7] Sorbonne Université/CNRS, FR2424, ABiMS, Station Biologique, 29680, Roscoff, France
[8] CHU Nantes, Inserm UMS 016, CNRS UMS 3556, SFR Santé, Bird, Université de Nantes, 44000, Nantes, France
[9] CNRS UMR 7104, BiGEst, IGBMC, 1 rue Laurent Fries, 67404 Illkirch Cedex

Corresponding Author: lecorguille@sb-roscoff.fr, seilerj@igbmc.fr

Since 2018, the IFB (Institut Français de Bioinformatique) has deployed, in addition to the Cloud infrastructure, a central HPC resource: the IFB Core Cluster. It is implemented by a cross-functional team of about ten engineers from IFB regional platforms ("mutualised task force") who dedicate part of their time to this common project.

This resource is made of 4300 CPU-cores (Hyper-Threaded), 2 PB of storage and 9 GPU cards (Nvidia A100). The cluster size should double before the end of 2022. The IFB Core Cluster offers various services and tools to facilitate access to life sciences data processing. Users can have, free of charge, an account, a shared project space of 250 GB and 10k hours of CPU that can be extended on demand. They also have access to more than 450 pre-installed scientific tools, and public indexed databanks.

All these resources are available through different user interfaces: i) CLI (Command Line Interface) via the scheduler SLURM, ii) JupyterHub, a web interface that allow the usage of Notebooks and terminal, iii) RStudio Server and finally iv) UseGalaxy.fr. Finally, the IFB Core Cluster via UseGalaxy.fr is supporting several IFB flagship health projects: EMERGEN (genomic surveillance of SARS-CoV-2), ABRomics (research and monitoring project on AntiBiotic Resistance) through its API and workflow system.

Beyond that, the IFB Core Cluster, as other NNCR nodes, is managed using Infrastructure as a Code (IaC) which means that all the administration is based on Ansible recipes, code and Continuous Integration (CI). Thus, everyone can participate in its administration in complete safety.

The IFB Core Cluster was the starting point for the National Network of Computing Resources (NNCR), which today brings together more than 8 IFB clusters in a federation of tools, best practices and expertises. Through this shared infrastructure, the contributing engineers from all IFB regional platforms were able to test and validate together new solutions that they then deployed on their own infrastructures. Most of these developments have in common the fact that they were born within specific regional platforms, were then refactorized, generalized and deployed on the IFB Core Cluster to finally be redistributed more widely to other regional platforms.

Automatization of Quality Data Workflows at a Genomic Platform: the GeT-PlaGe solution

Jules SABBAN[1], Eden DARNIGE[1], Romain THERVILLE[1], Abdias Archimede PATIPE[1], Céline VANDECASTEELE[1], Céline NOIROT[2], Christophe KLOPP[2], Christine GASPIN[2], Denis MILAN[1], Cécile DONNADIEU[1], Gérald SALIN[1] and Claire KUCHLY[1]

[1] INRAE US 1426 GeT-PlaGe Genotoul, Chemin de Borde Rouge, 31326, Castanet-Tolosan, France

[2] INRAE MIAT UR875 Plateforme Bioinformatique Genotoul, Chemin de Borde Rouge, 31326, Castanet-Tolosan, France

Corresponding Author: claire.kuchly@inrae.fr

Next Generation Sequencing (NGS) technologies, including Illumina, PacBio and Nanopore, have become fundamental in modern genomic, transcriptomic and epigenetic applications. With a steadily growing demand, the challenge lies in simultaneously sequencing multiple samples of various types. In order to fully benefit from the high-throughput production of NGS technologies, an advanced level of automation must be achieved from library preparation to data quality analysis. At the INRAE GeT-PlaGe facility, we are currently implementing a flexible LIMS, developed at *Institut de Génomique*, that allows to completely automate our data quality control workflow and other time-consuming tasks. With the development of pipelines using the widely-adopted and scalable Nexflow framework [1], we have completely rebuilt and optimized our quality control workflow. In the poster, we will describe the methodology that we used to design this workflow.

Acknowledgments

We thank the GenoToul bioinformatics facility for their support in computing resources and data storage, and for helping us learning how to develop using the Nextflow framework.

We thank the development and production management team of Institut de Génomique (CEA) for their support and developing the NGL LIMS. https://github.com/institut-de-genomique/NGL

References

1. Di Tommaso, P., Chatzou, M., Floden, E. W., Barja, P. P., Palumbo, E., & Notredame, C. (2017). Nextflow enables reproducible computational workflows. Nature Biotechnology, 35(4), 316–319.doi:10.1038/nbt.3820

A new bioinformatics pipeline for the analysis of hepatitis B virus transcriptome by Nanopore sequencing coupled to 5'RACE

Xavier Grand[1,2], Doohyun Kim[2], Delphine Bousquet[2], Guillaume Giraud[2], Cyril Bourgeois[1], Fabien Zoulim[2] and Barbara Testoni[2]

[1] INSERM U1052, CNRS UMR-5286, Cancer Research Center of Lyon (CRCL), 69008, Lyon, France.

[2] Laboratoire de Biologie et Modelisation de la Cellule, Universite de Lyon, CNRS UMR 5239, INSERM U1210, Ecole Normale Superieure de Lyon, Universite Claude Bernard Lyon 1, F-69007 Lyon, France.

Corresponding author: xavier.grand@ens-lyon.fr

Background: The Hepatitis B Virus has a circular DNA genome expressing 8 major viral RNAs (preC, pgRNA, preS1, preS2, S, L-HBx, HBx and S-HBx) and 22 splice variants (20 deriving from pg, 2 from preS1). The organization of the HBV genome is highly condensed and all transcripts are to various degrees subsets of each other. Indeed, all HBV transcripts share the same 3' end and, thus, the HBx sequence constitutes the 3' end of every viral transcript. Moreover, preC and pgRNA are longer than the viral DNA genome and thus are characterized by a 100bp redundancy at their 3' ends identical to their 5'end. Due to these constraints, HBV RNAs are indistinguishable by classic approaches like quantitative real time PCR and also by the short-reads transcriptome sequencing technologies, where in a large majority of cases, it is impossible to assign a short-read to a specific transcript. Our laboratory has previously described a Full length 5'RACE method which allows the qualitative assessment of all HBV transcripts according to their size, except for spliced variants, which require a Sanger sequencing step to be specifically recognized [1]. Therefore, it is so far impossible to specifically quantify each HBV transcript and splice variant produced during viral infection. Our aim is to propose a bioinformatics method to evaluate the expression of each individual HBV RNA and Splice Variants.

Methods: Following RNA extraction from HBV infected human hepatocyte cells or sera of chronically infected patients, we subjected the 5'RACE amplified HBV transcripts to Oxford Nanopore Technologies Long-read sequencing. We designed a new, HBV specific reference transcriptome to improve the mapping of reads and to reduce overestimation of longest viral RNAs and underestimation of shorter ones. Hence, we developed a dedicated graphical representation of the quantification of "start positions" of mapped reads on the HBV genome to obtain a global view of HBV transcriptome expression pattern. Moreover, we implemented the count of splice junctions to refine the evaluation of splice variants expression.

Results: We generated simulated and experimental data sets to improve and test our method. We showed that the quantification of canonical viral RNAs (not spliced) is accurate, with 0,1 to 10,9% of error. However, due to the high similarity of splice variants, in terms of sequence and position on the genome, the evaluation of their expression was still highly biased. Indeed, since only the use of specific splice junction or combination of junctions permits to distinguish these variants, we implemented the count of splice junctions usage to determine the expression of each splice variants.

Conclusion: We present here a bioinformatics method, based on different mapping approaches and analyses, that permits to accurately evaluate the expression of the known HBV transcriptome, coupling a 5'RACE amplification of full-length viral RNAs and long-read sequencing technology. The method is being implemented in a Nextflow pipeline to be published and released to the research community. This method will greatly help in improving the comprehension of HBV biology and will assist in the evaluation of the antiviral activity of newly developed anti-HBV compounds.

References

[1] Bernd Stadelmayer, Audrey Diederichs, Fleur Chapus, Michel Rivoire, Gregory Neveu, Antoine Alam, Laurent Fraisse, Kara Carter, Barbara Testoni, and Fabien Zoulim. Full-length 5'race identifies all major hbv transcripts in hbv-infected hepatocytes and patient serum. *Journal of Hepatology*, 73(1):40–51, 2020.

MaDMP4ls, or how to better manage bioinformatics projects with MY

Konogan Bourhy[1] and Olivier Collin[2]
[1] CNRS, Institut Français de Bioinformatique, Plateforme Genouest, Campus de Beaulieu, 35042 Rennes, France
[2] Plate-forme GenOuest Univ Rennes, Inria, CNRS, IRISA F-35000 Rennes

Corresponding author: konogan.bourhy@irisa.fr

In a context of exponential growth in the volumes of data generated, due to the increasing availability of new technologies and the metadata to accompany them, data management is taking on an increasingly important role in the research world. To address this global issue, the FAIR Principles [1] were established to provide guidelines for promoting the machine actionability of data so that it can be easily processed by computing tools. These FAIR principles have developed and taken hold in the various scientific communities, particularly in Biology. Like many other bioinformatics platforms, Genouest has witnessed this evolution, and has seen its role change from that of a simple calculation provider to that of a data management player, which has led to the creation of a data management and work environment, Cesgo [https://www.cesgo.org/fr/].

To facilitate the application of FAIR principles within the IFB infrastructure, we looked at DMPs (Data Management Plans) and the benefits they could bring to all stakeholders in a research project. This document, intended for the project funder, was initially created to describe, at the design stage of the research project, the different data to be produced, the processing methods, as well as where and how to access them once the project was completed. Despite the beneficial introspective work it represents, this document is still seen as an administrative burden with no real benefit to the researcher, which led to the launch of the MaDMP4LS project. This project is a collaboration between the IFB, the Genouest platform and the Opidor team of INIST and has two facets. Firstly, it aims to transform the OPIDoR DMP model into an actionable machine model, so that various systems and project actors can exploit the information contained in it. Secondly, the aim is to integrate the DMPs written and hosted on DMP OPIDoR [https://dmp.opidor.fr/] into the IFB infrastructure, via MY, the account and project management tool developed within the Genouest unit (IRISA, Rennes).

In concrete terms, we will show how, after having written his DMP on the OPIDoR DMP tool, the researcher will be able, via the MY tool, to request storage resources from the Genouest platform, simply by providing access to his DMP. This will enable the needs of a project to be defined automatically using information extracted from the DMP. The DMP will remain linked to the project, thus maintaining access to essential project information.

In the next phases of the project, we will work on automating this project request directly from the Opidor DMP interface, as well as allowing local modifications of the project to be sent back to the DMP so that the latter remains up to date without requiring manual intervention by the researcher. This project is also part of a larger project, MUDIS4LS, an ambitious IFB project that aims to promote open science in the life sciences by helping researchers apply the FAIR principles, and the maDMPs are a key element towards that goal.

References

[1] Mark D. Wilkinson, Michel Dumontier, IJsbrand Jan Aalbersberg, et al. The fair guiding principles for scientific data management and stewardship. *Scientific Data*, 3(1):160018, Mar 2016.

Expé-1point5 : une expérimentation nationale unique pour faciliter et étudier la transition des labos de recherche vers une réduction de leurs émissions de gaz à effet de serre

Sophie SCHBATH[1] et l'équipe Expérimentation de LABOS-1POINT5

[1] Université Paris-Saclay, INRAE, MaIAGE, 78350, Jouy-en-Josas, France

Corresponding Author: sophie.schbath@inrae.fr

Dans le contexte d'un changement climatique dont les effets se font de plus en plus prégnants, la question de la mesure et de la réduction de l'empreinte environnementale de la recherche académique – et en particulier son empreinte carbone – se pose. Le collectif Labos 1point5 [1] s'est saisi en 2019 de cette question qui devient un véritable objet de recherche. Constitué en GDR depuis 2021, Labos 1point5 vise tout d'abord à mesurer et à comprendre les déterminants de cette empreinte carbone à l'échelle de la recherche française, dans une approche très originale à cette échelle. L'outil GES 1point5 [2] développé dans cette perspective en constitue un outil de diagnostic et un premier pas vers l'action.

Pour stimuler, faciliter la mise en mouvement des laboratoires vers la réduction de leurs émissions de gaz à effet de serre (GES) et étudier les facteurs susceptibles de l'accélérer ou de la freiner, l'équipe Expérimentation propose différents dispositifs (de sensibilisation, financiers ou règlementaires) qui sont actuellement explorés dans une vingtaine de laboratoires participant à une phase pilote. Dans le cadre de cette « Expé-1point5 », un volet accompagnement a été développé afin de faciliter (i) l'appropriation de ces dispositifs par les laboratoires (Kit-1point5, qui présente l'ensemble de ces dispositifs), (ii) les discussions qui doivent y avoir lieu en vue de choisir collectivement un objectif de réduction, une trajectoire de réduction et un ou des dispositifs à explorer et (iii) les échanges de bonnes pratiques entre les laboratoires (développement d'une plateforme collaborative). Un volet recherche se constitue également pour étudier les processus en jeu lors de cette transition dans les laboratoires et pour explorer cette expérimentation sous différents angles disciplinaires, comme dans une approche plus intégrée. Seront également élaborés des indicateurs (i) de sobriété de la recherche, (ii) de qualité de la recherche sous contrainte climatique et (iii) de qualité de vie au travail, dans une tentative de réappropriation de cet enjeu climatique par la communauté de recherche.

Cette expérimentation, unique en son genre à cette échelle nationale, entrera dans une phase de déploiement au début 2023. Tous les laboratoires qui souhaiteront intégrer ce réseau de laboratoires en transition seront alors les bienvenus pour participer à cette dynamique de réappropriation de ces grands enjeux par la communauté de recherche. Ainsi, après avoir alerté, la communauté des personnels de la recherche et de l'enseignement supérieur se met en mouvement et espère entrainer avec elle d'autres pans de la société civile.

Le laboratoire MaIAGE fait partie des 20 labos pilotes pour cette expérimentation et vient de voter un objectif de 40% de réduction de ses émissions à l'horizon 2030. Ce poster sera l'occasion de faire un retour d'expérience, entre autres, sur la conduite des discussions au sein du labo et sur les actions et dispositifs de réduction expérimentés.

References

1. https://labos1point5
2. Jérôme Mariette, Odile Blanchard, Olivier Berné, et al. and Tamara Ben-Ari. An open-source tool to assess the carbon footprint of research. *bioRxiv* 2021.01.14.426384 doi: https://doi.org/10.1101/2021.01.14.426384

KATY european consortium: supporting the AI revolution in precision oncology

Florian Jeanneret[1], Pauline Bazelle[1], Etienne Bardet[1], Solène Mauger[1], Odile Filhol[1], Laurent Guyon[1], Delphine Pflieger[1], Stéphane Gazut[2], Jean-François Deleuze[3], Christophe Battail[1] and KATY consortium[4]

[1] Laboratoire Biologie et Biotechnologies pour la Santé, IRIG, UMR 1292 INSERM-CEA-UGA, Univ. Grenoble Alpes, 38000 Grenoble, France.

[2] Université Paris-Saclay, CEA, List, Palaiseau, France.

3 Centre National de Recherche en Génomique Humaine, CEA, Université Paris-Saclay, 91057 Evry, France.

[4] https://katy-project.eu, Européen Unions's Horizon 2020 research ans innovation programme, Grant agreement No 101017453

Corresponding Author: christophe.battail@cea.fr

Cancer research has been transformed in recent years by the considerable increase in omics data accumulated in public databases through the use of high-throughput technologies for profiling patient cohorts (1, 2). The challenge now is to translate the molecular characteristics of a patient's tumor into an appropriate therapeutic choice applicable in the clinic. Among the arsenal of anti-tumor treatments, targeted therapies and immunotherapies are now considered relevant therapeutic options used in first line for several types of cancer. However, a fairly large proportion of patients do not respond to these treatments or quickly develop resistance. The implementation of personalized medicine in our hospitals aims to help doctors better diagnose and treat their patients, by adapting the therapeutic choice to the molecular and cellular characteristics of each patient's tumor.

Artificial intelligence (AI) algorithms are powerful tools on which to rely to advance precision medicine. However, these AI-based approaches are often "black boxes" regarding the reasons leading to a decision, penalizing their use in the clinic. The KATY project seeks to build a precision medicine platform based on AI systems. It will be hosted on a European computing infrastructure and accessible by several European hospitals. This platform will not only be efficient, but above all transparent when it comes to the molecular, cellular and clinical evidence underlying the recommandation of drug treatments adapted to each patient. Clinicians will be able to trust, evaluate and effectively use this AI system in their daily work.

The platform implemented by the KATY consortium will be built around two main components: a distributed knowledge graph (DKG) and a collection of explainable artificial intelligence predictors (XAIPs). While the DKG is an intelligent repository that stores vast multi-omics patient information, as well as scientific information, the XAIPs will enrich the DKG and enable understandable personalized medicine decisions. This platform will be prototyped to predict the response of patients with kidney cancer to targeted therapies and immunotherapies.

References

[1] International Cancer Genome Consortium, Hudson, T. J., Anderson, W., Artez, A., Barker, A. D., Bell, C., Bernabé, R. R., Bhan, M. K., Calvo, F., Eerola, I., Gerhard, D. S., Guttmacher, A., Guyer, M., Hemsley, F. M., Jennings, J. L., Kerr, D., Klatt, P., Kolar, P., Kusada, J., Lane, D. P., … Yang, H. (2010). International network of cancer genome projects. *Nature*, *464*(7291), 993–998. https://doi.org/10.1038/nature08987

[2] Freeberg, M. A., Fromont, L. A., D'Altri, T., Romero, A. F., Ciges, J. I., Jene, A., Kerry, G., Moldes, M., Ariosa, R., Bahena, S., Barrowdale, D., Barbero, M. C., Fernandez-Orth, D., Garcia-Linares, C., Garcia-Rios, E., Haziza, F., Juhasz, B., Llobet, O. M., Milla, G., Mohan, A., … Rambla, J. (2022). The European Genome-phenome Archive in 2021. *Nucleic acids research*, *50*(D1), D980–D987. https://doi.org/10.1093/nar/gkab1059

ABRomics - a digital platform on antimicrobial resistance to store, integrate, analyze and share multi-omics data

Pierre MARIN[1,2], Julie LAO[1,3], Kenzo-Hugo HILLION[4], Nadia GOUE[2], [consortium ABRomics][5], Philippe GLASER[6] and Claudine MEDIGUE[1,7]

1. CNRS, Institut Français de Bioinformatique, IFB-core, UAR 3601, Évry, France
2. AuBi platform, Mésocentre, Université Clermont-Auvergne, 63170 AUBIERE, France
3. Université Paris Cité and Univ Sorbonne Paris Nord, Inserm, IAME, 75018 Paris, France
4. Faculté des Sciences et Techniques, Université de Nantes, 44000 Nantes, France
5. https://ppr-antibioresistance.inserm.fr/fr/projets-actions-soutenus/action-2-resultats-des-aap-structurants-du-ppr-antibioresistance-3-projets-retenus/#abromicspf
6. Institut Pasteur, Unité EERA, CNRS UMR604, 75015 Paris, France
7. CNRS UMR8030, Univ Evry-Val-d'Essonne, CEA, Genoscope, LABGeM, Evry, France

Corresponding Authors: pierre.marin@france-bioinformatique.fr, claudine.medigue@france-bioinformatique.fr

Antibiotic resistance (ABR) is a major public health issue prioritized for mitigation by major international institutions, especially regarding the emergence and the global dissemination of multidrug resistant (MDR) isolates and of antibiotic resistance genes (ARGs) carried by mobile genetic elements. They are also transmitted between humans, animals and the environment, without borders. The evolution of ABR is a complex process with multiple selective forces in different environments.

Genome sequences, which contain all the genetic information of an organism, can be used for molecular typing purposes with the highest resolution, the identification of ARGs and their genetic supports as well as mutations leading to a decrease in antibiotic susceptibility. Combined with epidemiological information, bacterial Whole Genome Sequencing (WGS) can enable tracking transmissions of outbreaks and identifying a source of contamination. Reinforcing the sharing of high-quality sequence data for diagnostic and epidemiological applications, together with interoperable and curated metadata, which can be integrated with other omics data, is a key requirement for understanding the complexity of spatiotemporal patterns of pathogen and ARGs transmission between compartments.

Today, if more systematic genome sequencing and bioinformatics analysis can partially address such major issues, data sharing and comparison across centers, together with standardization of analytical workflows remain major bottlenecks. The ABRomics project aims to develop a secure One Health, online platform to make bacterial infectious disease (meta)genomics data and their associated clinical and epidemiological metadata accessible to a meta-network of researchers including epidemiologists, clinical microbiologists and the wider research community. It should provide the diverse communities with a user-friendly platform to store, share and analyze genomic information together with their metadata. It will also serve as a data brokering hub to ease the submission of data sequences into international repositories.

The platform will meet two main objectives:

1. Establish a repository of structured, interoperable, standardized, and well-annotated multi-omics microbiology data with tailored mathematical and bioinformatics tools that can be used to answer specific research questions related to ABR.

2. Establish a shared platform to facilitate retrospective and prospective surveillance of ABRs in human and veterinary medicine, including environmental and food isolates, to enable near-real-time surveillance of pathogen transmission and outbreaks with actionable results for public health authorities. FAIR (Findable, Accessible, Interoperable and Reusable) data management procedures will enable retrospective studies.

We will present here the ongoing development of the three main modules of the ABRomics platform architecture: (i) *ABRomics-BIOINFO*: an IT infrastructure with high capacity of data storage and data analysis, accessible to users under defined rules and offering a general software environment, both HPC and Cloud environments, (ii) *ABRomics-DB*: an integrated multi-omics microbiological databases for the Human-Animal-Environment sectors, and (iii) *ABRomics-WF*: standard tools and pipelines for (semi) automatic analysis of NGS data from pathogenic strains.

The ABRomics project is led by the Institut Français de Bioinformatique (IFB), the Institut Pasteur and is made up of a consortium of 45 specialized teams belonging to the main French research organizations.

Montpellier Omics Days: An annual bioinformatics and biostatistics conference organised by Bioinformatics students

Yascim KAMEL[1], Nassif SAAB[1], Marie MILLE[1], Corentin MARCO[1], Fabien KON-SUN-TACK[1], Carla HEREDIA[1], Quentin BOUVIER[1], Bioinformatics Master's students[1] and Statistics and Data Science Course's students[1]

Université de Montpellier, 30, Faculté des Sciences de Montpellier, Place E. Bataillon, 34095 Montpellier, France

Corresponding author: yascim.kamel@etu.umontpellier.fr

Montpellier Omics Days (MOD) [1] is an event about the "omics" sciences, which refers to the combination of multiple disciplines for a better understanding of biological subjects. The pooling of disciplines for a more in-depth study tends to generate large amounts of data, a phenomenon accentuated by the newly emerging, more efficient and more affordable sequencing technologies. Analysing and storing all this data are becoming more and more challenging. To solve these problems, scientists try to implement information technologies, which include the creation of statistical methods and various tools capable of analysing and visualising different kinds of data. Consequently, "bioinformatics" came into being as an interdisciplinary field that combines knowledge from multiple domains such as biology and informatics. It is particularly through events like MOD that new problems and solutions in the field of bioinformatics are brought to the attention of biologists, bioinformaticians and biostatisticians. In February of 2022, MOD celebrated a decade of yearly held international conferences, organised by second year Bioinformatics Master's students at the University of Montpellier in France and students studying Statistics and Data Science at its Faculty of Sciences.

With over 500 subscriptions this year, the 10th edition brought together members of the scientific community worldwide. This poster presents what went into the organisation of a free-access event, the tasks entrusted to the students and the decisions they had to make in light of the constantly evolving laws and the restrictions brought about by the COVID-19 pandemic. Inquiring funds, finding the speakers and planning the event, from setting up the program to managing the subscriptions, are some of the tasks entrusted to the organisers. Divided into six sub-categories, i.e. management, finance, logistics, website & publicity, program and workshops, tasks are assigned to small groups of students. Having to quickly adjust to new regulations, the event was finally held online on January 26 in the form of a webinar, and the conferences were centered around the theme "BIODIVERSITY TECHNOLOGIES", chosen by the students among four subjects via anonymous voting. This edition welcomed seven speakers: Lucie Bittner, Benjamin Linard, Kateryna Makova, Nicolas Gilbert, Stéphanie Bocs, Simon Rio and Nika Abdollahi. Subjects ranging from metagenomics to immunology were presented to the spectators who then had the opportunity to engage with the speakers though Q&A sessions. The event concluded with a word from pioneer organisers of MOD and a closing speech thanking all the partakers and looking forward to next year's edition.

Acknowledgements

We would like to thank the supervisors of both masters, namely Sèverine BÉRARD and Annie CHATEAU, head of the Bioinformatics Master's, and Mrs. Elodie BRUNEL-PICCINNI, head of the Statistics and Data Science Course. We would also like to thank Anna-Sophie FISTON-LAVIER, former head of the Bioinformatics Master's, for her time, her guidance and and the logistical help she has provided.

Special thanks to our long-term partners, the University of Montpellier, the Faculty of Sciences, the Faculty of Medicine, the department of Informatics, the department of Mathematics, the Master of Bioinformatics and IMT Mines Alès, and to the parterns of this year's edition, GDR, CNRS, FSDIE, DRED, LabEx NUMEV, MUSE, JeBiF, SFBI and the newly-formed ÉBiM association.

References

[1] https://www.montpellier-omics-days.fr/.

Green-BIM: a study to make young bioinformaticians aware of the carbon footprint of bioinformatics

Emma Corre[1], Valentine Gilbart[1], Delphine Lanselle[1], Marie Lahaye[1], Matthias Lorthiois[1], Manea Meslin[1], Marie Vesselle[1], Marie Gspann[1,2] and Hélène Dauchel[1,3]

[1] Master's Degree of Bioinformatics, University of Rouen Normandy, Place E. Blondel 76821, Mont Saint Aignan, France

[2] Department of languages and communication, University of Rouen Normandy, Place E. Blondel 76821, Mont Saint Aignan, France

[3] LITIS UR 4108- FR NormaSTIC CNRS 3638, Team TIBS, University of Rouen Normandy, CURIB, 25, rue Tesnière, 76821 Mont-Saint-Aignan, France

Corresponding Author: helene.dauchel@univ-rouen.fr

Since 2006, the University of Rouen Normandy (URN) has been committed to a sustainable development and quality of life at work approach. It has resulted in the implementation of management practices of the university: selective sorting of waste, energy management, differentiated management of green spaces, travel plan, etc. As part of its strategy for social responsibility and sustainable development (DD&RS label in French "Développement Durable et Responsabilité Sociétale"), URN is now part of a more integrated approach including its fundamental missions of research and education.

In this context, the Master of Bioinformatics of the URN is committed to raise awareness among its students, future actors and professional decision-makers, of the environmental impacts and challenges of digital technologies in terms of sustainable development (green computing or "Green-IT"). Indeed, due to the analysis of large biological data, bioinformatics research and its applications require the extensive usage of large-scale computational infrastructure. However, data centers or clouds contribute to the increase of energy consumption and the chain of consequences such as the increase of greenhouse gas, the impact air quality and finally the global warming and imbalance of ecosystems. Recently, researchers have been studying the impact of bioinformatics algorithms and computational strategies [1, 2]. The authors quantified the environmental costs of numerous bioinformatics tools and commonly analyses, such as genome scaffolding, genome and metagenome assembly, metagenome classification, RNA-Seq pipelines especially read alignment, genome-wide association analysis, phylogenetics or molecular docking.

Here, we present a first work of sensitizing young professionals in bioinformatics to the challenges and good practices in this topic. First, we determined their weekly carbon consumption during their apprenticeship period, representing different cases of computation practice, tools and bioinformatics applications (metagenomics, machine learning, transcriptomics, comparative genomics, phylogenetics, genome annotation). Their carbon consumption was calculated by estimating the energy draw of the algorithms and the carbon intensity of producing this energy at a given geographical location (in kilograms of CO2 equivalent units, kgCO2e), using the Green Algorithms model and online tool (www.green-algorithms.org) [1, 2]. Secondly, the important dispersion of results was explored by detailing the major features of their practice such as runtimes, used memory or used CPU/GPU or the location of their resources. Finally, based on "The Carbon Footprint of bioinformatics" article [2], we summarize some advice to reduce these values and to improve the environmental impact for a greener bioinformatics.

Acknowledgements

The students were supported by the CEA, the CNRS, the INRAE, the Pasteur Institute and the companies that host them during their apprenticeship during the Bioinformatics master's degree at URN.

References

1. L. Lannelongue, J. Grealey, and M. Inouye, « Green Algorithms: Quantifying the Carbon Footprint of Computation », Advanced Science, vol. 8, no 12, p. 2100707, 2021, doi: 10.1002/advs.202100707.
2. J. Grealey et al., « The Carbon Footprint of Bioinformatics », Molecular Biology and Evolution, vol. 39, no 3, p. msac034, mars 2022, doi: 10.1093/molbev/msac034.

JeBiF - Association for the Young Bioinformaticians of France

Xavier Bussell[1], Emma Corre[1], Slim El Khiari[1], Mathias Galati[1], Klaus von Grafenstein[1], Victor Gretzinger[1] and Sarah Guinchard[1]
JeBiF, France

Corresponding author: contact@jebif.fr

JeBiF, which stands for "Jeunes Bio-Informaticiens de France" [Young Bioinformaticians of France], is a french non-profit organisation created in 2008. This association is the French branch of a bigger network handled by the International Society for Computational Biology Student Council (ISCB-SC). There are 26 other local branches. Such local branches are called Regional Student Groups (RSG), hence the full denomination of the association: RSG France – JeBiF.

The main goal is to promote the development of the next generation of bioinformaticians. In order to reach this goal, we mainly provide networking opportunities and career advice. We are also actively advertising computational biology and bioinformatics to general audience by the mean of science-popularisation events. In the poster, we present with more details the different activities developed by RSG France – JeBiF. In particular:

— JeBiF-Pub: every month, in different cities, JeBiF sympathisers are invited to meet at a bar for a drink (paused for now due to sanitary constraints).

— Table Ouverte en Bioinfo (TOBi): every month in a bar of Paris, a professional bioinfor- matician is invited to present their studies and career. Due to the sanitary constraints, TOBi have been replaced by TOBirtuelles which happen online and are opened to everyone, and not only Parisians.

— Table Ronde: once a year with the volunteering Master of Bioinformatics, several alumni of the Master are invited to present their path since they finished the Master. This year, due to the sanitary constraints, we replaced these local events by one big event organised with the support of the human resources team of the Pasteur Institute dedicated to career development and support for scientists (Mission Accueil, Accompagnement et Suivi des Carrières des Chercheurs – MAASQ). We received more than a hundred attendees.

— JeBiF@JOBIM: annual workshop with scientific presentations, open-table with various subjects, and flash-talks for people who have been accepted at JOBIM and would like to advertise their poster.

— Fête de la science and Pint of Science: two yearly events of science popularisation. JeBiF volunteers are encourage to create and animate activities, or give a talk.

The events proposed by RSG France – JeBiF are opened to everyone. The adhesion, which is free of charge, gives access to the mailing list and allows you to vote at the general assembly. It is also a way to quantify our impact. In particular, it gives us more weight in our funding applications. More funding allow us to maintain the association in long term, to propose more events and of larger scale. To date, JeBiF has 111 adherents. But its actions rely on the participation of its volunteers. We are always happy to meet new faces, and there is space for everybody to develop their ideas. Do not hesitate to join us if you would be part of the adventure!

C

D

E

F

G

H

I

J

K

L

M

N

O

P

Q

R

S

T

U

V

W

Y

Z


sfbi


cnrs
GDR
Groupement
de recherche
BIM Bioinformatique moléculaire


IFB
INSTITUT FRANÇAIS DE BIOINFORMATIQUE


INRAE


Inria


GDR
GÉNOMIQUE
ENVIRONNEMENTALE


UNIVERSITÉ DE
RENNES 1


Oxford
NANOPORE
Technologies


illumina


Hewlett Packard
Enterprise


Biogenouest


Ville
Métropole


FONDATION
RENNES 1
Progresser Innover Entreprendre