JOBIM 2022

05>08JUIL

Université de Rennes 1 | RENNES

Proceedings

Keynotes
Contributed talks
Mini-Symposia

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, Nathalie 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

R Ville Métropole

FONDATION RENNES 1
Progresser Innover Entreprendre

Table des matières

Keynotes

Nextflow and NF-Core : growing your analysis with the right data sprinkler

Cédric NOTREDAME

Centre for Genomic Regulation (CRG), The Barcelona Institute
for Science and Technology, Barcelona, Spain

Invited speaker: cedric.notredame@crg.eu

Pipelines are everywhere these days, and their rapidly growing interconnection presents us with increasingly complex challenges. Scientific reproducibility was the first issue to be pointed out, but the social relevance of analytic pipelines means they will soon impact our everyday lives at various levels. Reproducibility and readability are much more than mere technical details in a medical setup. In this talk, I will present the genesis of Nextflow and how it grew from an internal lab tool up to becoming one of the work-horses of global genomics analysis. I will provide a simple overview of the underlying algorithm and what distinguishes it from alternative solutions. I will also discuss the impact of production software development in a research lab and how to deal with this. I will conclude with a few examples of massive scale-up going on in the lab and the importance of interoperable genomic pipelines for data integration in collective projects such as BovReg and EuroFAANG that bring together the farmed animals genomics community. I will wrap up with our future plans for pipeline integration, including NF-Benchmark, our ongoing project for dynamic biological benchmarking.

Keynote

Swim in the data tsunami

Pierre PETERLONGO

Univ Rennes, CNRS, Inria, IRISA-UMR 6074, F-35000 Rennes, France

Invited speaker: pierre.peterlongo@inria.fr

In a few years we have witnessed an incredible change in the genomic landscape. Sequencing has become a matter of a few hours of experiment for a low price. As a result, our community is flooded with data.

In this context, many analysis methods have become routine. One can think for example of mapping sequencing data to good reference genomes to determine simple variants such as SNPs or short indels. However, these methods become inefficient in many situations such as the analysis of non-model species for which the reference is poor or non-existent, or the analysis of complex metagenomic data, or more globally to decompartmentalize the analyses, usually performed at the scale of one or a few datasets.

In this talk, I will present various methods of analysis of sequencing data, seen through the prism of breaking down the barriers between the projects, for a better exploitation of the wealth of information offered by the democratization of sequencing.

Analyzing the evolution of the microbiome using metabarcoding data

Hélène MORLON

IBENS, Ecole Normale Supérieure, UMR 8197 CNRS, 75005, Paris, France

Invited speaker: helene.morlon@bio.ens.psl.eu

Phylogenetic analyses are central to study evolutionary dynamics, such as past dynamics of speciation and extinction and co-evolutionary dynamics. While such analyses are widespread in well-characterized groups of species, such as animals and plants, they are much more challenging in groups for which diversity is mostly known through environmental DNA techniques, as is the case of most microbial groups. I will present several approaches to addressing the difficulty of carrying phylogenetic analyses using metabarcoding data. In the case of diversification analyses, I will show that at least some results can be robust to high phylogenetic uncertainty, and that this uncertainty can be reduced by combining metabarcoding data with reference sequences. In the case of the (co)evolution of host associated microbiota, I will present new models for detecting vertical transmission as well as covariations between microbial taxa abundances through time. I will illustrate the talk with examples from the oceanic plankton, arbuscular mycorrhizal fungi, and the gut microbiota.

Keynote

A replication crisis in methodological computational research?

Anne-Laure BOULESTEIX

Institute for Medical Information Processing, Biometry and Epidemiology,
Ludwig-Maximilians-University, 81377, Munich, Germany

Invited speaker: boulesteix@ibe.med.uni-muenchen.de

Statisticians are often keen to analyze the statistical aspects of the so-called “replication crisis”. They condemn fishing expeditions and publication bias across empirical scientific fields applying statistical methods. But what about good practice issues in their own - methodological - research, i.e. research considering statistical (or more generally, computational) methods as research objects? When developing and evaluating new statistical methods and data analysis tools, do statisticians and data scientists adhere to the good practice principles they promote in fields which apply statistics and data science? I argue that methodological researchers should make substantial efforts to address what may be called the replication crisis in the context of methodological research in statistics and data science, in particular by trying to avoid bias in comparison studies based on simulated or real data. I will discuss topics such as publication bias, the design and necessity of neutral comparison studies and the importance of appropriate reporting and research synthesis in the context of methodological computational research by drawing an analogy with clinical research.

Transposable elements are a source of innovation and variability in our response to infection

Guillaume BOURQUE[1,2]

[1] Department of Human Genetics, McGill University, Montréal, Québec, H3A 0G1, Canada
[2] Canadian Center for Computational Genomics, McGill University, Montréal, Québec, H3A 0G1, Canada

Invited speaker: guil.bourque@mcgill.ca

Transposable elements (TEs) have expanded the binding repertoire of many transcription factors and, through this process, have been co-opted in different transcriptional networks. In particular, TEs have contributed new regulatory sequences to immune-related genes. However, the extent to which TEs have shaped the evolution of the mammal immunity against pathogens remains poorly understood. In this presentation, I will discuss the results from a new study looking at epigenetic changes following influenza infection in monocyte-derived macrophages extracted from 35 individuals of different ancestry. Specifically, I will show that different TEs families are associated with regions of shared, but also variable, accessibility across individuals upon infection. I will also demonstrate that TEs can predict a significant portion of the heterogeneity that we observe in the response to infection. Finally, I will present new bioinformatics methods that make use of graph genomes to analyse epigenomic data. Such graphs can reveal signal that would have been missed using a standard reference genome.

Exploring interactomes in 3D under the lens of coevolution

Raphaël GUÉROIS

Université Paris-Saclay, CEA, CNRS, Institute for Integrative Biology of the Cell (I2BC), 91198,Gif-sur-Yvette, France

Invited speaker: raphael.guerois@cea.fr

Predicting protein-protein interactions and characterizing their structural organization provides essential information to elucidate the molecular mechanisms underlying cross-talk between cellular pathways. Exploiting coevolution information has emerged over the last years as a major strategy for predicting the mode of recognition between proteins. Our recent methodological developments [1,2] have provided key insights to unravel the regulatory subtleties that might exist between pairs of interacting proteins [3,4]. The advent of machine learning in the field, culminating in the Alphafold breakthrough, is pushing the boundaries from analyses of binary to multiple interactions. We will discuss about strategies to further integrate predictions with experimental assessment and further unravel the complexity of the molecular logic behind large networks of proteins interacting together.

References

1. C. Quignot, P. Granger, P. Chacón, R. Guerois, and J. Andreani. Atomic-level evolutionary information improves protein-protein interface scoring. Bioinformatics, Apr 2021.
2. C. Quignot, G. Postic, H. Bret, J. Rey, P. Granger, S. Murail, P. Chacón, J. Andreani, P. Tufféry, and R. 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 Res, 49(W1):W277–W284, 07 2021.
3. A. Pyatnitskaya, J. Andreani, R. Guérois, A. De Muyt, and V. Borde. The Zip4 protein directly couples meiotic crossover formation to synaptonemal complex assembly. Genes Dev, 36(1-2):53–69, 01 2022.
4. A. Acharya, K. Kasaciunaite, M. Göse, V. Kissling, R. Guérois, R. Seidel, and P. Cejka. Distinct RPA domains promote recruitment and the helicase-nuclease activities of Dna2. Nat Commun, 12(1):6521, 11 2021.

Peer Community In: A free alternative to evaluate, validate (and publish) preprints

Denis BOURGUET

INRAE, UMR CBGP (INRAE-IRD-CIRAD-Montpellier SupAgro), Campus International de Baillarguet, 34988, Montferrier sur Lez Cedex, France

Invited speaker: denis.bourguet@inrae.fr

The Peer Community in (PCI,https://peercommunityin.org) project offers an alternative to the current system of publication - which is particularly expensive and not transparent. PCI is a non-profit scientific organization building communities of researchers handling the evaluation (through peer-review) and recommendiation of preprints in their scientific field. Each PCI is a group of several hundred recommenders playing the role of editors who recommend preprints based on peer-reviews to make them complete, reliable and citable articles, without the need for publication in 'traditional' journals (although the authors can submit their recommended preprints afterwards). Evaluations and recommendations by a PCI are free of charge. When a recommender decides to recommend a preprint, they write a recommendation text that is published along with all the editorial correspondence (reviews, recommender's decisions, authors' replies) on the PCI website. The preprint itself is not published by PCI: it remains on the preprint server where it has been posted by the authors and can therefore be submitted to a journal or published in "Peer Community Journal", an open access diamond journal launched in 2021. The first Peer Community In was started in 2017: Peer Community in Evolutionary Biology and there are now 15 thematic PCIs, including PCI Mathematical anc Computational Biology. More than 1,700 scientists from around the world have already joined as PCI recommenders. The PCI initiative won the 2020 LIBER award for library innovation of the European League of Research Libraries.

Contributed talks

Session 1, Amphi A

Workflows, reproducibility & open science I

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

Keywords Galaxy, FAIR data analysis, reproducibility, computing infrastructure

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 [2], Jupyter [3], RShiny [4], ...). Galaxy also allows to manage scientific data by easily sharing and publishing scientific results and workflows, with customisable reporting and visualisations. As Galaxy keeps track of each user action, and thanks to its technical architecture, reproducibility of each result is guaranteed, following the FAIR principles. This web portal makes computational biology accessible to scientists without requiring computer programming skills.

The French Institute of Bioinformatics (IFB) [5] 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. Building on this infrastructure, a national Galaxy instance, UseGalaxy.fr (https://usegalaxy.fr), 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.

In addition to the global UseGalaxy.fr portal, five thematic subdomains have been set up (Workflow4Metabolomics [6], Covid19, ProteoRE [7], Metabarcoding, Ecology). These sub-domains allow users to find a more focused choice of specific tools, while benefiting from the same technical infrastructure.

Following the international UseGalaxy.* [8] guidelines, administration and development of UseGalaxy.fr is based on a IaC (Infrastructure as Code) model: every aspect of the portal is managed using Ansible roles and playbooks (developed collaboratively with the international Galaxy community [9]), deposited on public GitLab repositories [10]. As such, in a collaborative and open-to-contributions manner, UseGalaxy.fr is continuously being improved and adapted to the needs of the community by implementing new features or making available new software and data collections.

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 this talk, we will propose a general introduction to UseGalaxy.fr and to the latest recent developments in the Galaxy portal (workflows, interactive tools, ...). We will also present how UseGalaxy.fr can be used as a complete training platform, by using training material from the Galaxy Training Network [11] (a collection of tutorials for users, developers and admins), and TIaaS [12] (Training Infrastructure as a Service), a new service deployed recently on UseGalaxy.fr. TIaaS is a tool to assist in the organisation of training sessions: it allows to easily 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, we will focus on several IFB flagship projects making use of UseGalaxy.fr, notably for data processing, integration and sharing needs, but also for the implementation of specific tools. In particular:

- Emergen [13]: project for the genomic surveillance of SARS-CoV-2 in France
- Abromics [14]: research and monitoring project on antibiotic resistance
- Dedicated Galaxy subdomains (Workflow4Metabolomics, Covid19, ProteoRE, Metabarcoding, Ecology)

References

[1] Enis Afgan, Dannon Baker, Bérénice Batut, Marius van den Beek, Dave Bouvier, Martin Čech, John Chilton, Dave Clements, Nate Coraor, Björn Grüning, Aysam Guerler, Jennifer Hillman-Jackson, Vahid Jalili, Helena Rasche, Nicola Soranzo, Jeremy Goecks, James Taylor, Anton Nekrutenko, and Daniel Blankenberg. 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] RStudio Team (2020). RStudio: Integrated Development for R. RStudio, PBC, Boston, MA URL http://www.rstudio.com/.

[3] Kluyver, T. et al., 2016. Jupyter Notebooks – a publishing format for reproducible computational workflows. In F. Loizides & B. Schmidt, eds. Positioning and Power in Academic Publishing: Players, Agents and Agendas. pp. 87–90.

[4] Winston Chang, Joe Cheng, JJ Allaire, Carson Sievert, Barret Schloerke, Yihui Xie, Jeff Allen, Jonathan McPherson, Alan Dipert and Barbara Borges (2021). shiny: Web Application Framework for R. R package version 1.7.1. https://CRAN.R-project.org/package=shiny

[5] http://www.france-bioinformatique.fr/

[6] Franck Giacomoni, Gildas Le Corguillé, Misharl Monsoor, Marion Landi, Pierre Pericard, Mélanie Pétéra, Christophe Duperier, Marie Tremblay-Franco, Jean-François Martin, Daniel Jacob, Sophie Goulitquer, Etienne A. Thévenot, Christophe Caron, Workflow4Metabolomics: a collaborative research infrastructure for computational metabolomics, Bioinformatics, Volume 31, Issue 9, 1 May 2015, Pages 1493–1495, https://doi.org/10.1093/bioinformatics/btu813

[7] Combes, F., Loux, V., & Vandenbrouck, Y. (2021). GO enrichment analysis for Differential Proteomics Using ProteoRE. In D. Cecconi (Ed.), Proteomics Data Analysis (pp. 179–196). Springer US. https://doi.org/10.1007/978-1-0716-1641-3_11

[8] The Galaxy Community. Galaxy Community Update. BCC2020. https://vimeo.com/440168116

[9] https://galaxyproject.org/

[10] https://gitlab.com/ifb-elixirfr/usegalaxy-fr

[11] Batut B, Hiltemann S, Bagnacani A, Baker D, Bhardwaj V, Blank C, Bretaudeau A, Brillet-Guéguen L, Čech M, Chilton J, Clements D, Doppelt-Azeroual O, Erxleben A, Freeberg MA, Gladman S, Hoogstrate Y, Hotz HR, Houwaart T, Jagtap P, Larivière D, Le Corguillé G, Manke T, Mareuil F, Ramírez F, Ryan D, Sigloch FC, Soranzo N, Wolff J, Videm P, Wolfien M, Wubuli A, Yusuf D; Galaxy Training Network, Taylor J, Backofen R, Nekrutenko A, Grüning B. Community-Driven Data Analysis Training for Biology. Cell Syst. 2018 Jun 27;6(6):752-758.e1. doi: 10.1016/j.cels.2018.05.012. PMID: 29953864; PMCID: PMC6296361.

[12] Helena Rasche, Björn Grüning Training Infrastructure as a Service. bioRxiv 2020.08.23.263509; doi: https://doi.org/10.1101/2020.08.23.263509

[13] https://emergen-db.france-bioinformatique.fr/

[14] https://www.france-bioinformatique.fr/actualites/abromics-une-plateforme-numerique-sur-la-resistance-antimicrobienne-pour-stocker-integrer-analyser-et-partager-des-donnees-multi-omiques/

EMERGEN-BioInfo : The digital platform for the French SARS-CoV-2 genomic surveillance and research program

Thomas Denecker[1], 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 Médigue[1,9], Gildas Le Corguillé[3,1], David Salgado[1,10], Jacques van Helden[1,4]

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: thomas.denecker@france-bioinformatique.fr

Abstract *We present EMERGEN-Bioinfo, the digital platform to collect, process, manage and divulgate viral sequences and non-sensitive metadata, developed in the context of EMERGEN, the French plan for COVID-19 genomic surveillance and research. The bioinformatics platform relies on different components to manage all the steps from raw sequence collection to deposition in international repositories. This includes: (1) specific storage spaces for each one of the 60 teams of the consortium; (2) a data lake gathering all sequences (raw, mapped, consensus genomes, aligned genomic and peptidic sequences); (3) system-level workflows to handle the data flow trough all the components of the platform; (4) a covid-19 specific domain of the national Galaxy server (covid19.usegalaxy.fr); (5) EMERGEN-DB, a database to store and manage non-sensitive metadata and genomic consensus sequences ; (6) data brokering services to facilitate metadata management and curation, submission to international repositories (GISAID[1,2,3] and ENA[4]) and follow-up of their acceptance status. The EMERGEN-Bioinfo platform is complemented by a high-security digital platform (EMERGEN-HDS) certified for Health Data Storage, which will enable researchers to pair EMERGEN data with patient data from different sources (national COVID-19 and healthcare databases). All the software resources developed for this projects will be accessible under an open license, and re-usable for other national projects (e.g. ABRomics multi-omics platform for surveillance and research on antimicrobial resistance) or international cooperation (e.g. sharing with partners of the European bioinformatics infrastructure ELIXIR).*

Keywords SARS-CoV-2; COVID-19; genomic surveillance; health data; EMERGEN-Bioinfo.

1. Introduction

In January 2021, the French ministries of Health (MSS) and Research (MESRI) launched a national plan for SARS-CoV-2 genomic surveillance and research, which aims at monitoring the evolution of the COVID-19 in France, at detecting new variants, and at enabling the integration of viral genomic and patient health data. The project is co-led by Santé publique France (SpF), the French national public health agency, and ANRS-Maladies infectieuses émergentes (ANRS-MIE), the French Agency for Research on AIDS, Viral Hepatitis and Emerging infectious diseases (ANRS-MIE) and relies on a wide network of public and private partners (Figure 1). Samples are collected via a wide national network encompassing more than 5,000 sampling sites, in community-based private laboratories and public healthcare facilities. Viral genome sequencing relies on a consortium involving 54 sequencing laboratories (8 high-capacity sequencing platforms including the National Reference Laboratory for respiratory viruses, 39 hospital laboratories of the ANRS-MIE network, and 7 other private laboratories). The data is collected, handled and accessed via two digital platforms enabling to analyze respectively the non-sensitive data (EMERGEN-Bioinfo, operated by the French Institute of Bioinformatics) and the health-related data (EMERGEN-HDS, operated by Inserm). EMERGEN data is used for two complementary purposes : surveillance (managed by SPF and the Centre National de Référence des virus respiratoires) and research (managed by the ANRS|MIE). The sequences and non-sensitive metadata are published in two international repositories (GISAID and EBI-ENA). Access to extended metadata can be obtained for research or surveillance projects by submitting projects to the EMERGEN data access committee.

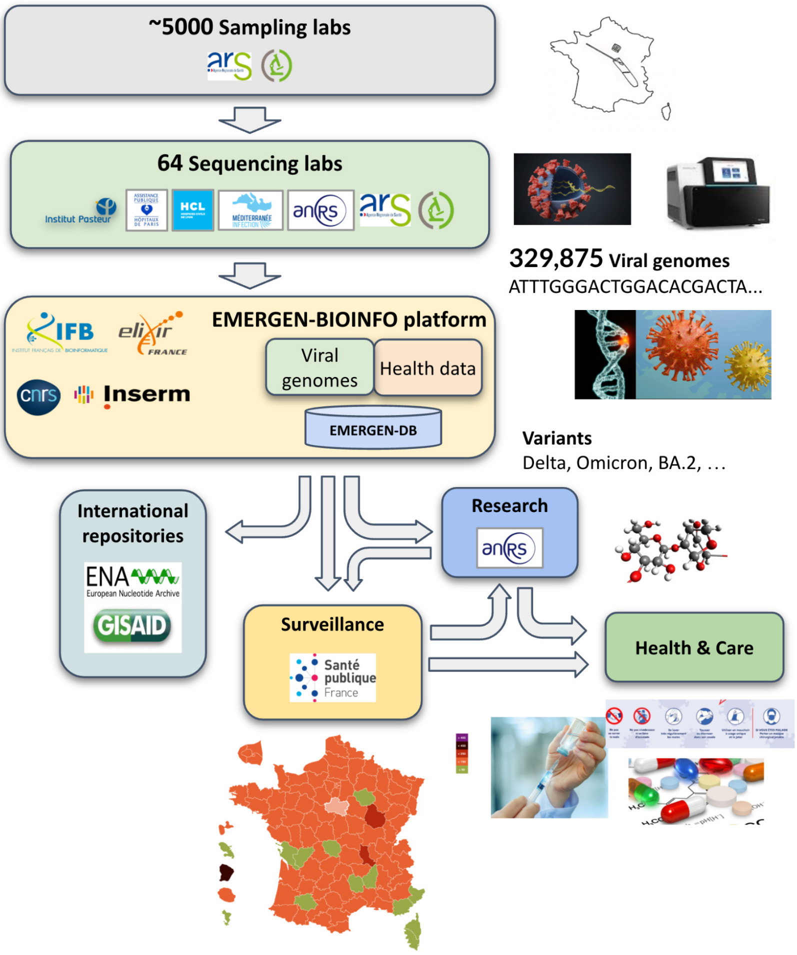


Figure 1. Organization of the EMERGEN plan.

The IFB, in collaboration with Inserm, is in charge of developing the digital platform and bioinformatics services enabling to collect, organize, analyze and divulgate the sequences and metadata produced by the EMERGEN consortium. The digital platform is splitted into two major components: ***EMERGEN-BioInfo***, operated by IFB, to ensure the treatment of viral sequences and non-sensitive metadata, and ***EMERGEN-HDS***, operated by Inserm, to enable the pairing between EMERGEN data and patient data from other sources (national COVID and health and care databases) in a high-security infrastructure certified for the hosting of health data (HDS, for Hébergement de Données de Santé).

2. EMERGEN data flow

EMERGEN-BioInfo platform handles three types of data collected from the sequencing labs (Figure 2). Firstly the raw genomic data used to run the workflows in the IFB cluster. Secondly, the consensus genomes and metadata that includes information about sampling, sequencing and variants which are stored and handled by EMERGEN-DB. Finally, the identifying metadata used either for surveillance in order to be paired with other COVID databases (SI-DEP, SI-VIC, VAC-SI) and also enrich the EMERGEN metadata, or used for research inside the Health Data Storage (HDS) certified space in an attempt to pair them with health data.

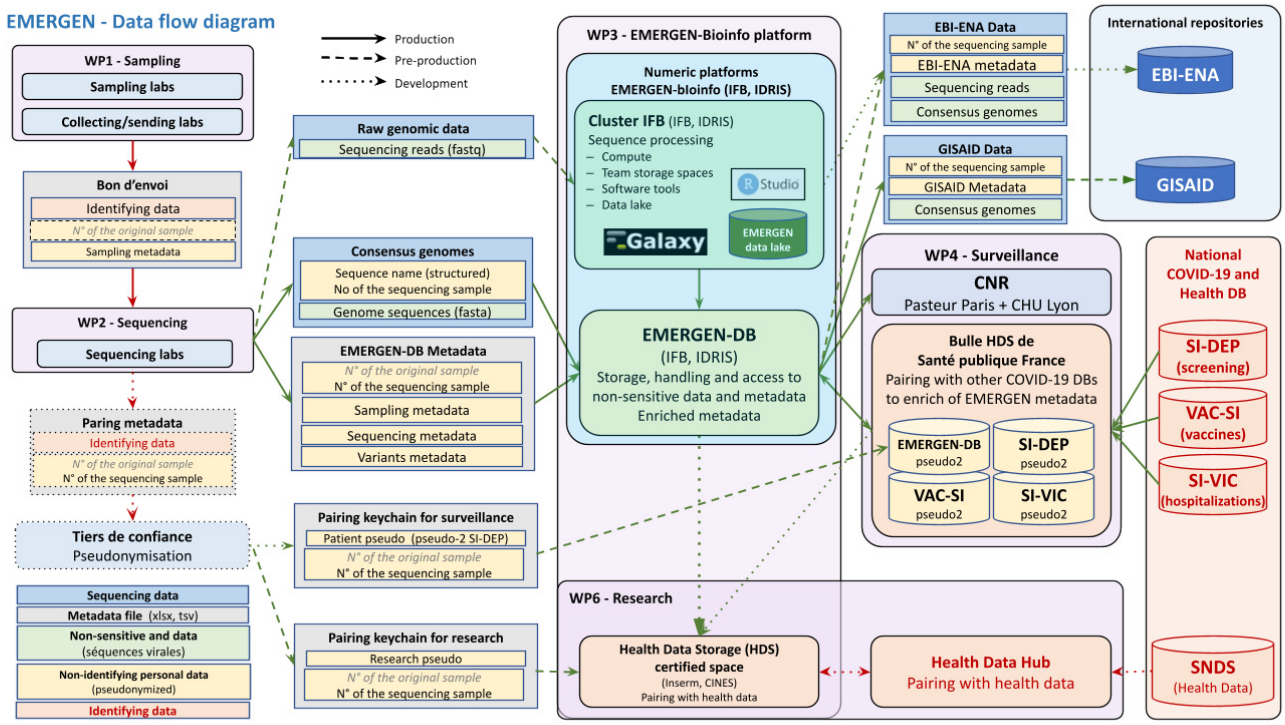


Figure 2. EMERGEN data flow.

3. Components of the EMERGEN-Bioinfo digital platform

We describe hereafter the different components of the EMERGEN-Bioinfo digital platform (Figure 3).

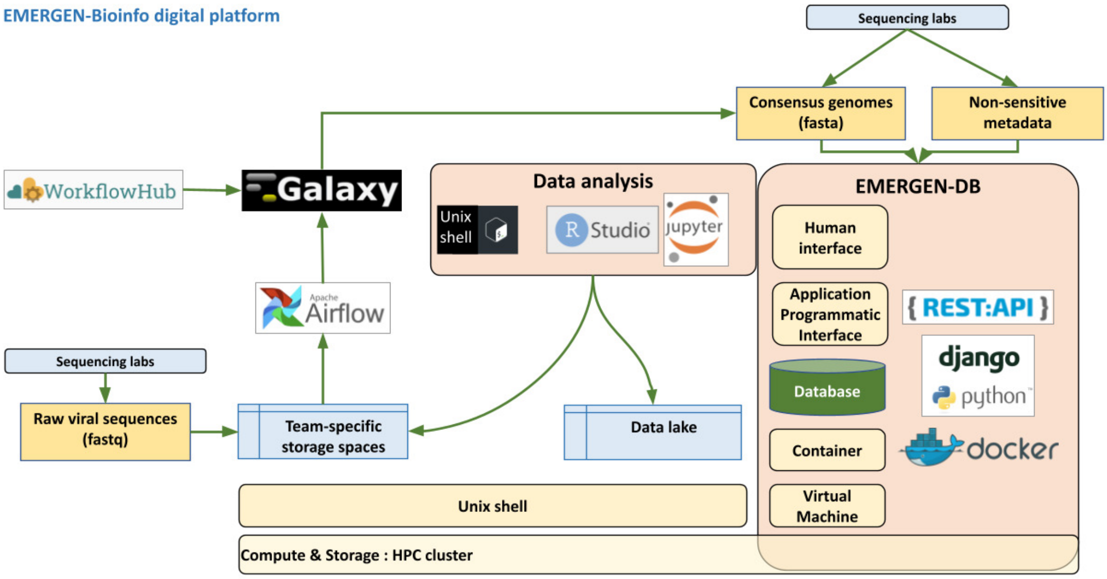


Figure 3. Components of the EMERGEN-Bioinfo digital platform.

3.1. Compute & storage HPC cluster

Maintained by the IFB, the HPC cluster provides a ready-to-use environment offering an extensive software catalog that covers all types of bioinformatics analyses. A specific project space has been created for each team of the consortium, enabling sequencing labs to deposit their raw and processed data. Each team-specific project space is also accessible by the project managers (IFB and SpF).

The IFB cluster also deploys two user-friendly analysis environments, RStudio (rstudio.cluster.france-bioinformatique.fr) and JupyterHub (jupyterhub.cluster.france-bioinformatique.fr), enabling statistical or bioinformatics analyses to be carried out on the data from individual and team project spaces. This software environment was, for instance, used to evaluate sequencing quality of specific laboratories and to study the evolution of Omicron variant contamination across the French territory.

3.2. Data lake

A data lake will allow people authorized by the access committee to consult all the viral sequences produced by the sequencing laboratories. and by the bioinformatics platform: raw sequences, aligned genomic and peptidic sequences. The data lake is updated regularly and automatically, and its purpose is to later aggregate most of the data/results produced by the IFB component of the EMERGEN team. It will be accessible via the Unix shell as well as the Web interfaced analysis environments (RStudio, JupyterHub).

3.3. System-level management of the data flow (Airflow)

An automatic scheduler based on the ETL ("Extract, Transform, Load") Airflow[5], handles with different automatons batches of raw data uploaded by the sequencing facilities. Its primary purpose is to analyze on the fly the fastq or the bam files by chaining several bash or Galaxy workflows: decontamination, analysis processed, reporting and publication. It also serves to automate, schedule and monitor other tasks such as variants reanalysis from the consensus sequences deposited to EMERGEN-DB.

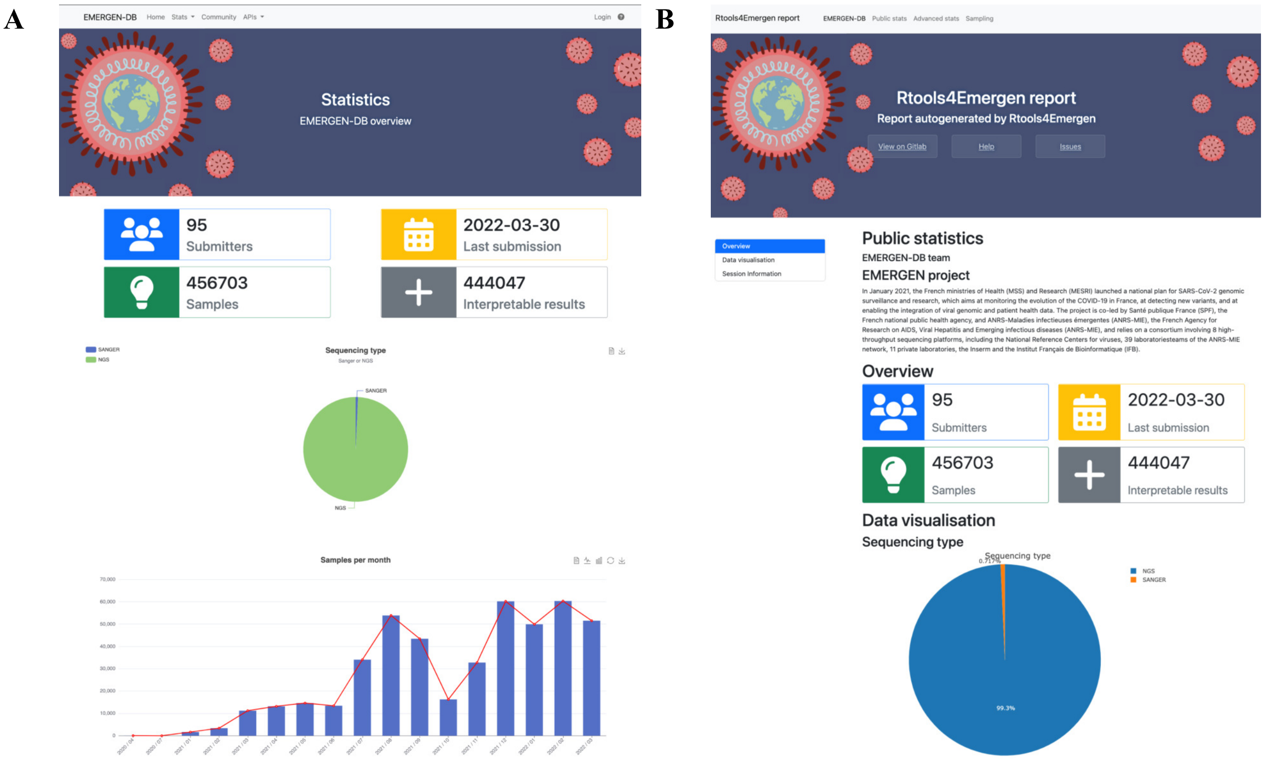
3.4. Analysis workflows (Galaxy)

A COVID-19 specific Galaxy[6] domain was deployed on the IFB core cluster (covid19.usegalaxy.fr), to run analysis workflows from the raw reads deposited on the IFB cluster by any sequencing laboratory contributing to the EMERGEN project. This Galaxy instance provide access to SARS-CoV-2 genomic and proteomic sequences; 7 workflows specifically designed to analyze SARS-CoV-2 NGS data, from assembly to variant analysis, produced by different sequencing technologies (Illumina, Nanopore), and maintained worldwide by the IWC (<https://github.com/iwc-workflows>);

3.5. EMERGEN-DB

EMERGEN-DB (<https://emergen-db.france-bioinformatique.fr/>) is the database that collects and organizes non-sensitive metadata (sample collection, sequencing method, …) and consensus SARS-CoV-2 genomic sequences. Developed under Django, it is equipped with both user-friendly and application programmatic interfaces (API), enabling users to upload data and query the databases manually or in batch. EMERGEN-DB also offers numerous tools to facilitate real-time monitoring of variants (alerts, etc.), data exploration (interactive pages), data visualization and data export (via the API).

*The R package **Rtools4emergen*** (<https://gitlab.com/ifb-elixirfr/covid19/rtools4emergen>) provides a collection of *R* functions enabling to query EMERGEN-DB via its API and to generate different types of graphical representations. It provides flexible support to generate reproducible scientific reports.

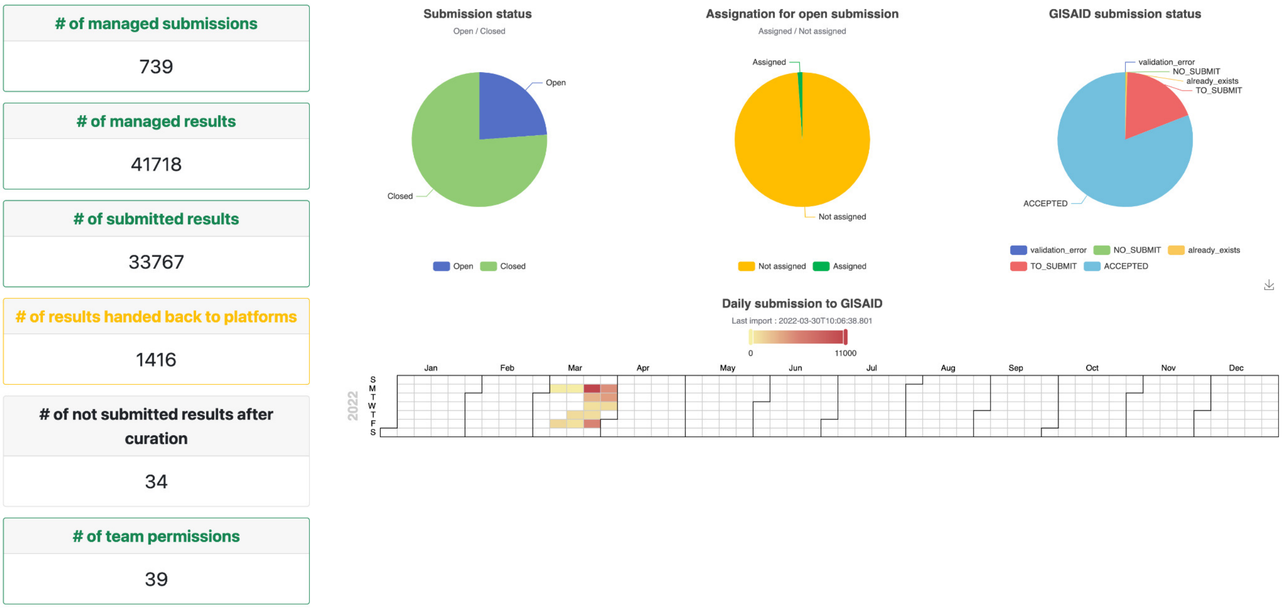


In January 2021, the French ministries of Health (MSS) and Research (MESRI) launched a national plan for SARS-CoV-2 genomic surveillance and research, which aims at monitoring the evolution of the COVID-19 in France, at detecting new variants, and at enabling the integration of viral genomic and patient health data. The project is co-led by Santé publique France (SPF), the French national public health agency, and ANRS-Maladies infectieuses émergentes (ANRS-MIE), the French Agency for Research on AIDS, Viral Hepatitis and Emerging infectious diseases (ANRS-MIE), and relies on a consortium involving 8 high-throughput sequencing platforms, including the National Reference Centers for viruses, 39 laboratoriesteams of the ANRS-MIE network, 11 private laboratories, the Inserm and the Institut Français de Bioinformatique (IFB).

Figure 4. Public statistics page from EMERGEN-DB website (A) and from the automatic report generated by Rtools4Emergen (B)

3.6. Support to data brokering

EMERGEN-DB also includes a service of data brokering to support the automated submission of genomic sequences to two international repositories : GISAID to ensure a rapid sharing of the consensus genomes and non-sensitive metadata (in production), and EBI-ENA to ensure an open access of the same data plus the raw sequences (in development). As of March 30th 2022, the EMERGEN-DB has collected metadata for 456,703 samples, which allowed the collection of 379,242 SARS-CoV-2 genomes and 444,047 variants coming from laboratories from all over France.



# of managed submissions

739

# of managed results

41718

# of submitted results

33767

# of results handed back to platforms

1416

# of not submitted results after curation

34

# of team permissions

39

Figure 5. Example of the monitoring dashboard (i.e. : GISAID dashboard March 30th, 2022)

4. Handling of personal and health-related data

Viral genomic sequences are not sensitive, but the metadata collected on the sampling platforms contains personal information about the patient (age, sex, region of residence) and medical status (sequencing indications e.g. vaccination, reinfection, immunodepression, return from foreing country…) which might indirectly lead to reidentification if not properly handled. To preserve patient privacy, the metadata submitted to international repositories is restricted to a carefully selected subset that would not enable reidentification.

Patient identities are however required to enable pairing between EMERGEN and the information gathered in different COVID-19 national databases (Figure 2): SI-VIC (hospitalization and death), SI-DEP (screening) and VAC-SI (vaccination), and for the national database of health and care (SNDS).

5. Conclusion and perspectives

The EMERGEN-Bioinfo digital platform is based on a wide range of computing resources developed by the international open source community and by the European infrastructure in bioinformatics ELIXIR.

Consistently with the open science standards and engagement, the EMERGEN-Bioinfo digital platform will ensure compliance with the FAIR principles.

- A Data Management Plan is under development.
- All the workflows developed by the team will be accessible on Workflowhub.
- The codes of the different components (***Rtools4emergen***, ***EMERGEN-Workflows***, ***EMERGEN-DB***) have their respective code versioned (GitLab).
- The ***EMERGEN-DB*** application is also distributed in a docker image autogenerated by continuous integration after verification of the quality of the code (test coverage set at 100% and code lint)
- A detailed metadata referential has been produced, which serves as framework to generate template files to facilitate the metadata curation and validation, and to transfer the different types of metadata (sampling, sequencing, pairing, export to international repositories) between the successive steps of the data flow. The metadata referential and template files will be available in Zenodo.

According with IFB commitments in the context of this pandemia, we are working closely with international community in particular ELIXIR[8], in the context of the CONVERGE WP9[9], and PHA4GE[10], to homogenize metadata description standards and investigate data exchange protocols such as GA4GH[11] ELIXIR Beacon[12] for SARS-CoV-2 data.

This system has been designed to provide a generic framework to deploy bioinformatics platforms for genomic surveillance and research, which will be reusable in other contexts. In particular, it will readily be used by the French multi-omics platform for surveillance and research about antimicrobial resistance (ABRomics). It might also be reused by other member nodes of the ELIXIR networks, enabling them to rapidly deploy a flexible and powerful bioinformatics platform, and to contribute collectively to its further developments.

Availability

The following web addresses were available on 23 March 2022.

EMERGEN-DB front-end	emergen-db.france-bioinformatique.fr
French COVID-19 Galaxy server	covid19.usegalaxy.fr
EMERGEN-DB code	gitlab.com/ifb-elixirfr/covid19/EMERGEN-DB
RTools4Bioinfo (R package)	gitlab.com/ifb-elixirfr/covid19/rtools4emergen

Acknowledgements

This work was supported by funding of Santé publique France, the French national public health agency ; “Enhancing Whole Genome Sequencing (WGS) and/or Reverse Transcription Polymerase Chain Reaction (RT-PCR) national infrastructures and capacities to respond to the COVID-19 pandemic in the European Union and European Economic Area” Grant Agreement ECDC/HERA/2021/007 ECD. 12221 ; the Institut Français de Bioinformatique (IFB), ANR-11-INBS-0013.

References

1. Shruti Khare, Céline Gurry, Lucas Freitas, Mark B Schultz, Gunter Bach, Amadou Diallo, Nancy Akite, Joses Ho, Raphael TC Lee, Winston Yeo, GISAID Core Curation Team, Sebastian Maurer-Stroh. GISAID’s Role in Pandemic Response[J]. China CDC Weekly, 2021, 3(49): 1049-1051. doi: 10.46234/ccdcw2021.255
2. Elbe, S. and Buckland-Merrett, G. (2017) Data, disease and diplomacy: GISAID’s innovative contribution to global health. Global Challenges, 1:33-46. doi:10.1002/gch2.1018 PMCID: 31565258
3. Shu, Y. and McCauley, J. (2017) GISAID: from vision to reality. EuroSurveillance, 22(13) doi:10.2807/1560-7917.ES.2017.22.13.30494 PMCID: PMC5388101
4. Leinonen R, Akhtar R, Birney E, Bower L, Cerdeno-Tárraga A, Cheng Y, Cleland I, Faruque N, Goodgame N, Gibson R, Hoad G, Jang M, Pakseresht N, Plaister S, Radhakrishnan R, Reddy K, Sobhany S, Ten Hoopen P, Vaughan R, Zalunin V, Cochrane G. The European Nucleotide Archive. Nucleic Acids Res. 2011 Jan;39(Database issue):D28-31. doi: 10.1093/nar/gkq967. Epub 2010 Oct 23. PMID: 20972220; PMCID: PMC3013801.
5. https://airflow.apache.org/
6. Enis Afgan, Dannon Baker, Bérénice Batut, Marius van den Beek, Dave Bouvier, Martin Čech, John Chilton, Dave Clements, Nate Coraor, Björn Grüning, Aysam Guerler, Jennifer Hillman-Jackson, Vahid Jalili, Helena Rasche, Nicola Soranzo, Jeremy Goecks, James Taylor, Anton Nekrutenko, and Daniel Blankenberg. 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
7. https://usegalaxy.fr/
8. Blomberg, N., Lauer, K.B. Connecting data, tools and people across Europe: ELIXIR’s response to the COVID-19 pandemic. Eur J Hum Genet 28, 719–723 (2020). DOI: doi.org/10.1038/s41431-020-0637-5

9. https://elixir-europe.org/about-us/how-funded/eu-projects/converge/wp9

10. Griffiths EJ, Timme RE, Mendes CI, Page AJ, Alikhan NF, Fornika D, Maguire F, Campos J, Park D, Olawoye IB, Oluniyi PE, Anderson D, Christoffels A, da Silva AG, Cameron R, Dooley D, Katz LS, Black A, Karsch-Mizrachi I, Barrett T, Johnston A, Connor TR, Nicholls SM, Witney AA, Tyson GH, Tausch SH, Raphenya AR, Alcock B, Aanensen DM, Hodcroft E, Hsiao WWL, Vasconcelos ATR, MacCannell DR. Future-proofing and maximizing the utility of metadata: The PHA4GE SARS-CoV-2 contextual data specification package. Gigascience. 2022 Feb 16;11:giac003. doi: 10.1093/gigascience/giac003. PMID: 35169842; PMCID: PMC8847733.

11. Rehm, Heidi L., et al. "GA4GH: International policies and standards for data sharing across genomic research and healthcare." Cell genomics 1.2 (2021): 100029. https://doi.org/10.1016/j.xgen.2021.100029

12. https://beacon-project.io

Session 1, Amphi B

Metagenomics & microbial ecosystems

Biosynthetic potential of the global ocean microbiome

Lucas PAOLI[1], Hans-Joachim RUSCHEWEYH[1*], Clarissa C. FORNERIS[2*], Florian HUBRICH[2*], Satria KAUTSAR[3], Agneya BHUSHAN[2], Alessandro LOTTI[2], Quentin CLAYSSEN[1], Guillem SALAZAR[1], Alessio MILANESE[1], Charlotte I. CARLSTRÖM[1], Chrysa PAPADOPOULOU[1], Daniel GEHRIG[1], Mikhail KARASIKOV[4,5,6], Harun MUSTAFA[4,5,6], Martin LARRALDE[7], Laura M. CARROLL[7], Pablo SÁNCHEZ[8], Ahmed A. ZAYED[9], Dylan R. CRONIN[9], Silvia G. ACINAS[8], Peer BORK[7,10,11], Chris BOWLER[12,13], Tom O. DELMONT[13,14], Josep M. GASOL[8], Alvar D. GOSSERT[15], André KAHLES[4,5,6], Matthew B. SULLIVAN[8,16], Patrick WINCKER[13,14], Georg ZELLER[7], Serina L. ROBINSON[2,17✉], Jörn PIEL[2✉], Shinichi SUNAGAWA[1✉]

[1] Department of Biology, Institute of Microbiology and Swiss Institute of Bioinformatics, ETH Zürich, Zürich 8093, Switzerland

[2] Department of Biology, Institute of Microbiology, ETH Zürich, Zürich 8093, Switzerland

[3] Bioinformatics Group, Wageningen University, 6708PB Wageningen, The Netherlands

[4] Department of Computer Science, ETH Zürich, Zurich 8092, Switzerland

[5] Biomedical Informatics Research, University Hospital Zurich, Zürich 8091, Switzerland

[6] Swiss Institute of Bioinformatics, Lausanne 1015, Switzerland

[7] Structural and Computational Biology Unit, European Molecular Biology Laboratory, Meyerhofstrasse 69117 Heidelberg, Germany

[8] Department of Marine Biology and Oceanography, Institute of Marine Sciences ICM-CSIC, Barcelona 08003, Spain

[9] Center of Microbiome Science, EMERGE Biology Integration Institute, Department of Microbiology, The Ohio State University, Columbus, Ohio 43210, USA

[10] Max Delbrück Centre for Molecular Medicine, Berlin 13125, Germany

[11] Department of Bioinformatics, Biocenter, University of Würzburg, Würzburg 97074, Germany

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

[13] Research Federation for the Study of Global Ocean Systems Ecology and Evolution, FR2022/Tara Oceans GOSEE, 75016 Paris, France

[14] Metabolic Genomics, Genoscope, Institut de Biologie François Jacob, CEA, CNRS, Univ Evry, Université Paris Saclay, 91000 Evry, France

[15] Department of Biology, Biomolecular NMR Spectroscopy Platform, ETH Zürich, Zürich 8093, Switzerland

[16] Department of Civil, Environmental and Geodetic Engineering, The Ohio State University, Columbus, Ohio 43210, USA

[17] Department of Environmental Microbiology, Eawag: Swiss Federal Institute of Aquatic Science and Technology, Dübendorf 8600, Switzerland.

[*] these authors contributed equally

[✉] corresponding authors

Corresponding authors: ssunagawa@ethz.ch; jpiel@ethz.ch; srobinson@ethz.ch

Paper Reference: **Paoli *et al*. Biosynthetic potential of the global ocean microbiome, *Nature*, In Press, https://doi.org/10.1038/s41586-022-04862-3 (*bioRxiv*, 2021, https://doi.org/10.1101/2021.03.24.436479)**

Introduction

Microbes drive global biogeochemical cycles, support food webs and underpin the health of animals and plants [1]. Their immense phylogenetic, metabolic and functional diversity represents a rich discovery potential for new taxa [2], enzymes and biochemical compounds, including natural products. In addition to their original functions, these natural products and the genetically encoded pathways for their production have led to numerous biotechnological and therapeutic applications [3]. The discovery of such pathways and compounds have largely been facilitated by studying cultivable microbes under laboratory conditions. However, taxonomic surveys of natural environments have revealed that the vast majority of microbial life has not yet been cultivated [4]. To overcome these limitations, technological advances over the past decade have enabled researchers to directly (*i.e.*, without prior cultivation) sequence pieces of microbial DNA from whole communities (metagenomics) or single cells. The possibility to assemble such pieces into larger genomic fragments and to reconstruct several metagenome assembled genomes (MAGs) or single amplified genomes (SAGs), respectively, now enables the genome-resolved exploration of microbiomes (*i.e.*, microbial communities and their contained genetic material in a given environment). However, despite recent attempts to establish genome-resolved explorations of various environmental and host-associated microbiomes [5], for the ocean, the largest ecosystem on Earth, over two-thirds of global metagenomic data cannot be associated with known species [5,6]. Thus, the biosynthetic potential of the ocean microbiome in general and its potential as a reservoir of new enzymology and natural products specifically remain largely underexplored.

Results and discussion

Here, we integrated ~10,000 microbial genomes from cultivated and single cells with >25,000 newly reconstructed draft genomes from >1,000 globally-distributed seawater samples to establish the Ocean Microbiomics Database (https://microbiomics.io/ocean/), which provides genomic context for over 60% of prokaryote-enriched marine metagenomic data. Leveraging this resource, we explored the diversity and novelty of biosynthetic gene clusters (BGCs) in the ocean and revealed ~40,000 putative BGCs. To account for inherent redundancy, we grouped BGCs predicted to encode for similar natural products into gene cluster families (GCFs). Out of ~7,000 GCFs, we found that over half were likely to encode for new compounds, many of which coming from previously unsuspected phylogenetic groups. Among these groups, we discovered a BGC-rich lineage (*Candidatus* Eudoremicrobiaceae) that belongs to an uncultivated bacterial phylum and includes some of the most biosynthetically diverse microbes in this environment. From these, we characterized two pathways and revealed cases of unusual bioactive compound structure and enzymology, respectively: (1) the phospeptin pathway encodes for a poly-phosphorylated peptide with low-micromolar protease inhibitory activity and (2) the pythonamide pathway produces a massive, hypermodified peptide and includes a new type of enzyme catalyzing the challenging yet biotechnologically relevant backbone *N*-methylation of the amide bond. These are among the most complex ribosomally synthesized and post-translationally modified peptides from distant and uncultivated bacterial phyla produced using synthetic biology methods to date. Together, this work illustrates how microbiomics-driven strategies enable prospecting for yet unknown enzymes and natural products in underexplored microbial groups and environments.

References

1. Cavicchioli, R. et al. Scientists' warning to humanity: microorganisms and climate change. *Nat. Rev. Microbiol.* 17, 569–586 (2019).
2. Hug, L. A. et al. A new view of the tree of life. *Nature Microbiology* vol. 1 (2016).
3. Newman, D. J. & Cragg, G. M. Natural Products as Sources of New Drugs over the Nearly Four Decades from 01/1981 to 09/2019. *J. Nat. Prod.* 83, 770–803 (2020).
4. Lloyd, K. G., Steen, A. D., Ladau, J., Yin, J. & Crosby, L. Phylogenetically Novel Uncultured Microbial Cells Dominate Earth Microbiomes. *mSystems* 3, (2018).
5. Nayfach, S. et al. A genomic catalog of Earth's microbiomes. *Nat. Biotechnol.* 1–11 (2020).
6. Pachiadaki, M. G. et al. Charting the Complexity of the Marine Microbiome through Single-Cell Genomics. *Cell* 179, 1623–1635.e11 (2019).

Session 1, Amphi C

Algorithms & data structures for sequences I

Optimal Scaffolding for Chloroplasts' Inverted Repeats

Victor Epain, Rumen Andonov and Dominique Lavenier
Univ. Rennes, Inria, IRISA, CNRS, F-35000 Rennes, France

Corresponding author: victor.epain@irisa.fr

Abstract *Scaffolding step in the genome assembly aims to determine the order and the orientation of a huge number of previously assembled genomic fractions (contigs/scaffolds). Here we introduce a particular case of this problem and denote it by Nested Inverted Fragments Scaffolding (NIFS). We formulate it as an optimisation problem in a particular kind of directed graph that we call Multiplied Doubled Contigs Graph (MDCG). Furthermore, we prove that the NIFS problem is NP-Complete. We also discuss how the chloroplast data have been generated by filtering the reads sequenced both from plants and chloroplasts. Moreover, we propose a graph structure to visualise the solution and to highlight the particularity of chloroplast's regions structure.*

Keywords Genome assembly, Inverted repeats, De Bruijn graph, Assembly graph, NP-Complete

1 Introduction

Sequencing an organism is extracting DNA molecules contained in its cells and getting it as sequences of ATGC letters corresponding to nucleotides. Until now, sequencing technologies cannot return one complete sequence for each DNA molecule in cells but fragmented copies of them, called reads. Furthermore, sequencing DNA molecules from plants cells does sequence not only the DNA of the plant organism, but also the organelles living in them. This paper focuses on how to assemble chloroplast genome, a plant organelle responsible for photosynthesis metabolism, which confers the ability to transform sunlight energy to chemical one. The knowledge of chloroplast genomes allows evolutionary analyses [1], (meta)-barcoding [2], and is useful for biophotovoltaic process development [3].

Given a set of reads, assembling them needs comparisons between their sequences in order to detect if they overlap or not. The first difficulty for assembly methods is provoked by genomes' repeats: reads from one region can overlap reads from a repeat of it, while they do not correspond to same genomic location. Another difficulty is yielded by the fact that reads are sequenced from the two complemented DNA strands with no distinction. *Inverted repeats* (one region is the reverse-complement of the other one) can produce repeat-induced overlaps and so can lead to assemble a mix of two strands regions. One more specific issue here is how to separate reads from plant and the ones from the chloroplast genome. Indeed, partitioning them is not so trivial as some plant's reads

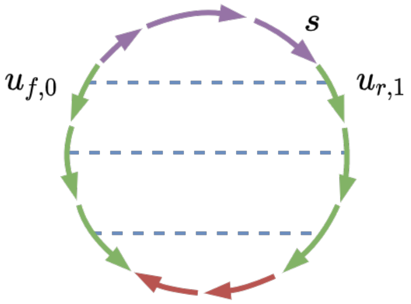


Fig. 1. Chloroplast genome's structure. Each arrow represents an oriented DNA sequence (contigs). Chloroplast's genome is circular thus it is a circular sequence of contigs, that begins at s and ends in s. Purple and red regions are unique, while green ones are *inverted repeats* (one is obtained by reversing and complementing the sequence of the other). For a given contig $u_{f,0}$ on the left-side green region, the contig in front of it (on the right-side green region) has the reverse orientation (so r). Because orientations are mutually exclusive, it is necessary to chose an occurrence that differs 0 (so 1). The couple $u_{f,0}, u_{r,1}$ is defined as inverted fragments. Each inverted fragments couple is linked by blue dashed line. Inverted repeats can be modelled as a sequence of nested inverted fragments (as illustrated with parallel blue dashed lines).

can overlap some chloroplast's reads. As a consequence, choosing repeat-induced or plant-chloroplast overlaps can lead to misassembled genome.

Dedicated chloroplast genomes assemblers have been already developed [4, 5] and have been compared in [6]. Exclusively based on the knowledge that chloroplast genome is circular and has two inverted repeats [7] (as illustrated in Fig. 1), we propose a dedicated scaffolding strategy that focuses on inverted repeats reconstruction [8]. This scaffolding approach is a linear programming formulation, which requires as input an assembly graph obtained from a pre-assembly step. Our method does not need any distance information between the reads and only uses genomic regions decomposition. Although our formulation is similar to the one proposed in [9] for RNA folding, here we deal with contigs instead of nucleotides and *a priori* without any knowledge concerning their order — which obviously increases the underlying challenges.

2 Method

Input data are reads that come from short reads sequencing technology, and can be part of plant's or/and chloroplasts' genomes. We also use chloroplast proteome data from a near plant species, and a well known gene sequence that is contained with a very high probability in one of the chloroplast's unique genomic regions, defined as the *seed gene*. An overview of the method is illustrated in Fig. 2. In the next section we briefly describe how the reads from the chloroplast's and the plant's genomes are separated, and how the remaining reads are assembled into contigs.

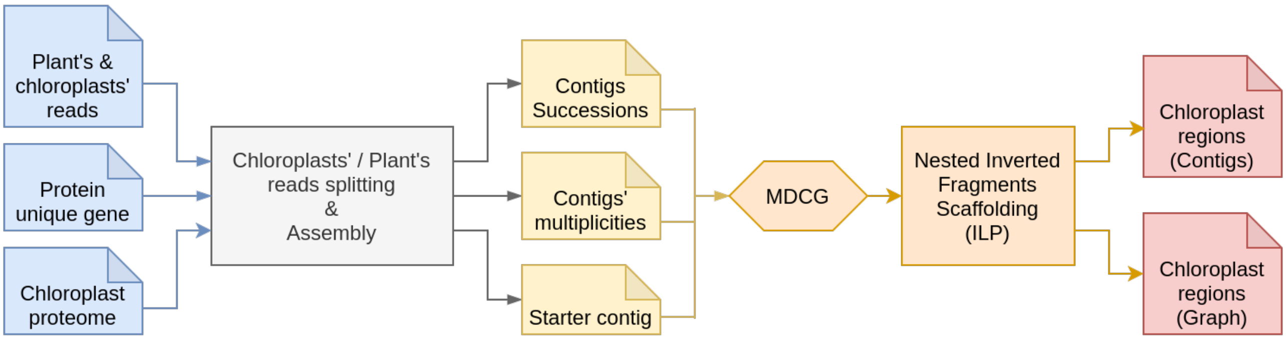

Plant's & chloroplasts' reads
Protein unique gene
Chloroplast proteome
Chloroplasts' / Plant's reads splitting & Assembly
Contigs Successions
Contigs' multiplicities
Starter contig
MDCG
Nested Inverted Fragments Scaffolding (ILP)
Chloroplast regions (Contigs)
Chloroplast regions (Graph)

Fig. 2. Method overview. The method can be cut in two main parts. First step is a driven chloroplast assembly with reads that come from both plant's genome and chloroplasts' genome. This assembly is a De Bruijn graph assembly approach. It outputs contigs' successions and an estimated multiplicity for each contig. These results are structured in a Multiplied Doubled Contigs Graph (MDCG). We finish the assembly thanks to an Integer Linear Programming (ILP) approach. Finally, we output a fewer number of larger contigs that correspond to chloroplast's genomic regions, and a scaffolding enriched assembly graph, as illustrated Fig. 4.

2.1 Data Filtering & Contigs Generation

Under the hypothesis that chloroplasts' genome are over-represented comparing to the plant's genome in the plant's cell, only a subset of order of million randomly chosen reads is considered. Thus, reads are hashed into k-mers, and we keep reads with a high k-mer coverage. This implies that the assembly is done using exclusively k-mers with a high coverage presumably sequenced from chloroplasts' genome (and repeats in the plant).

Reads from this subset are assembled into contigs with a De Bruijn graph (DBG) approach using the tool MINIA [10]. For each contig its *mean coverage* is computed according to the mean k-mer distribution over it. Then, the seed gene is mapped against contigs. The contig with the best alignment is defined as the *seed contig*, and its mean coverage becomes the *reference coverage*. Other contigs are mapped against translated chloroplast's proteome from a near plant's species thanks to TBLASTN tool [11]. Those which obtain a high alignment score are tagged as *contigs near to the proteome*.

Then we use MINDTHEGAP [12] to find k-mers paths in the DBG from tagged contigs to other tagged ones. These paths are defined as *links*. These links can pass through a limited number of non-tagged contigs: in that case, these contigs are selected to be in the final contigs set.

We define the contigs set $\mathcal{C}$ as the set containing near-to-the-proteome tagged contigs and those which are intermediate in paths between two tagged ones. For each contig we compute a *multiplicity value* as an approximated ratio between contig's estimated coverage and seed contig's coverage (the reference coverage). The seed contigs is expected to participate only once into the final assembly. Thus, a contig's multiplicity may be interpreted as an upper-bound of the number of times the associated contig can be met in the final assembly. Contigs and their attributes are illustrated in Tab. 1a.

If one contig is participating in the solution, its sequence is oriented: it can be either in its *forward* orientation (so the sequence does not change), or in its *reverse* orientation (the sequence is read in inverse reading and each nucleotide is complemented). For this reason links in the set $\mathcal{L}$ are succession relations between two oriented contigs. Tab. 1b shows links examples, where orientation attribute is indicated as 'f' (resp. 'r') for forward (resp. reverse) orientation.

c_{id}	c_{len}	c_{mult}
2	18914	2
3	19212	1
4	88398	1
5	7596	2

(a) Contigs data

c_{id}	c_{or}	d_{id}	d_{or}
2	r	3	r
2	r	3	f
2	f	5	r
4	f	5	f

(b) Links data

Tab. 1. Example of scaffolding input data. It contains two types of data: **(a)** Set $\mathcal{C}$ of contigs with their length (c_{len}) and multiplicity (c_{mult}); **(b)** Set of links $\mathcal{L}$. An identifier (c_{id}) and an orientation (c_{or}) is provided for each contig c. Concerning the orientation, 'f' and 'r' are resp. standing for *forward* and *reverse* contig's orientations. For example, first line means that from contig 2 in reverse orientation we can go to contig 3 in reverse orientation. For the sake of saving memory, for each link $(c, d) \in \mathcal{L}$, only one of the existing two links is reported in the table, the one with $c_{id} < d_{id}$.

2.2 Multiplied Doubled Contigs Graph

Building chloroplast genomic regions implies finding a sequence of linked oriented contigs, in the limit of multiplicity for the two orientations for each contig. Links between two multiplied (by their multiplicity) and oriented (forward or reverse) contigs can be structured in an oriented graph. Thus, we define $MDCG = (V, E)$ the *Multiplied Doubled Contigs Graph*, where V is the set of vertices and E is the set of edges. Even if in practice we do not multiply and double contigs and links data in memory, each vertex $v \in V$ is one occurrence of an oriented contig and each edge $(u, v) \in E$ corresponds to a link between occurrences of two oriented contigs. Let us show some MDCG properties:

- For each vertex $v \in V$, its reverse $\overline{v}$ is in the graph too. By definition, vertices' identifier remain the same $v_{id} = \overline{v}_{id}$, and same for occurrences $v_{occ} = \overline{v}_{occ}$.
- For each edge $e = (u, v) \in E$, its reverse $\overline{e} = (\overline{v}, \overline{u})$ is in the graph too.

Furthermore, we add what we define as *inverted fragments* data into MDCG. Each inverted fragments is an unoriented couple of vertices (i, j), such that they are two different occurrences of the same contig, but one is in forward orientation, while the other is in reverse orientation. These couples are candidates to be part of inverted repeats pairs. Fig. 1 shows an example where $i = u_{f,0}$ and $j = u_{r,1}$.

2.3 Nested Inverted Fragments Scaffolding

Thanks to the MDCG structure, building a sequence of linked oriented contigs is equivalent to finding an elementary path in MDCG. This is a necessary, but not a sufficient condition to correctly retrieve a circular genome with pairs of inverted repeats. Especially, no more that one from the vertices $v \in V$ and its reverse $\overline{v} \in V$ can be in the path. The seed contig is illustrated as a big black dot in Fig. 1 & 3 (seed vertex $s \in V$, in arbitrarily forward orientation). The circularity of the genome corresponds to a path that begins at s and ends in s.

Finding a path requires giving a position for each vertex participating in it. As illustrated in Fig. 1, two inverted fragments (i, j) and (k, l) in inverted repeats regions are nested *i.e.* if you draw a line from i to j, and another from k to l, then the two lines do not intersect. In order to know if two inverted fragments intersect, we must compare the positions of the associated vertices.

Finally, the goal of the Nested Inverted Fragments Scaffolding (NIFS) is to find a path from s to s, that passes through at most one of the orientations of multiplied contigs, and that maximises

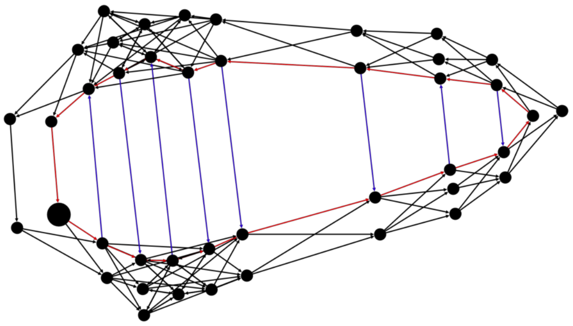


Fig. 3. An instance of *Multiplied Doubled Contigs Graph.* The input contigs were multiplied by their multiplicity number, then doubled according to two DNA strands. The obtained graph possesses 42 nodes and 130 edges. Nodes candidate to participate in inverted repeats have one of their reverse oriented versions linked by a blue edge. The solution (the assembled genome) is represented as a path in red. It passes through 8 adjacent inverted fragments that represent inverted repeats. It begins with the biggest node (a given starter, as illustrated in Fig. 1) and finishes in the same node since the chloroplast genomes are circular.

the number of nested inverted fragments. To solve this problem we give in [8] a linear programming formulation, such that vertices' position, their relative location, chosen edges and chosen inverted fragments are all integer or binary variables. We also demonstrate that the NIFS problem is NP-Complete. Moreover, the number of variables and constraints is polynomial according the vertices and edges number ($O(|V|^2 + |E|)$ for both) (*ibid.*).

3 Results

The mathematical formulation was implemented in Python3 using the PuLP package where we run Gurobi solver with academic licence. All the instances have been executed on a Linux laptop computer (32GB RAM, Intel® Core™ i7-10610U CPU @ 1.80GHz ×8).

Input data are contigs and links obtained as described in Section 2.1. When all links between contigs are provided, the computed solution enables to assemble the contigs and to output final sequences (fewer and longer contigs). As the genomes for these instances are known, it is possible to asses our solution. We used for this purpose Quast — a well known assembly evaluation tool [13]. All genome references are from NCBI database. Each instance has been run 10 times. Tab. 2 reports some solver run statistics and major Quast measures.

For each feasible instance, our scaffolding program outputs 3 contigs: two of them correspond to two unique regions known as long and short single copy (LSC & SSC), and the third one corresponds to one of the inverted repeats region (IRa or IRb). In addition, our tool outputs an assembly graph that visualises the connections between the three genomic regions. This assembly graph is enriched by reverse properties as illustrated in Fig. 4. Indeed, the graph in Fig. 4 shows that there are two possible circular sequences of genomic regions: $0 - 1_{f,0} - 2_f - 1_{r,1}$ and $0 - 1_{f,0} - 2_r$

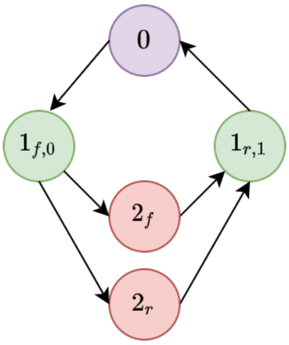


Fig. 4. Scaffolding final enriched assembly graph. Nodes correspond to genomic regions while edges indicate their neighbourhoods. Purple, green and red nodes correlate respectively with purple, green and red regions in Fig. 1. Two green nodes' label differ by their orientation 'f' and 'r' (one is the reverse complement sequence of the other). They also differ by their occurrence number (0 and 1). It simply means that their associated genomic regions are simultaneously present in the genome sequence (*i.e.* the green nodes both participate in the solution). On the opposite, red nodes just differ by their orientation: it implies that only one genomic region from one of these two nodes appears in the genome (one genomic region is the reverse complement of the other).

– $1_{r,1}$. These results are not contradictory because chloroplast genomes are multimeric: two conformations can simultaneously exist in the same plant's cell. In fact, one of the unique genomic region between the two inverted repeats (*e.g.* the red region in Fig. 1 & 4) exists in forward orientation in some genomes, and in reverse orientation in others. This multimeric property for chloroplast genomes has been studied in [14].

instance	**size**	$\|V\|$	$\|E\|$	**time**	**%gnm**	**NGA50**	**#mis**	**|mis|**	**mism**	**indels**
Aloysia citriodora	154699	42	74	.04	99.85	85509.0	0	0	3.88	10.36
Altheaea officinalis	159987	10	24	.01	99.91	87847.0	0	0	0.0	0.0
Amborella trichopoda	162686	106	168	.19	98.82	88782.0	0	0	162.35	36.08
Citrus limon	160101	174	358	.44	80.32	26819.0	2	102284	35.0	39.66
Coffea arabica	155188	134	222	.01	-	-	-	-	-	-
Dendrobium nobile	152018	36	54	.03	97.77	33841.0	2	105360	837.25	187.93
Digitalis lanata	153108	48	86	.08	83.15	83859.0	0	0	11.0	9.43

Tab. 2. Scaffolding results on chloroplast benchmark with reference genomes. Instance: species name; **size**: genome length in number of nucleotides (in base-pairs); $|V|$ and $|E|$ are respectively the number of vertices and edges; **time**: CPU solver time (in seconds); **%gnm**: percentage of genome length covered by the solution computed by our approach; **NGA50**: NG50 corrected of assembly errors; **#mis**: number of misassemblies in the proposed solution; **|mis|**: misassemblies total length (in bp); **mism**: number of alignment mismatches per 100 kbp; **indels**: number of alignment indels per 100 kbp.

Excluding three of the instances (Citrus limon, Dendrobium nobile and Coffea arabica) all other instances have been successfully solved. Citrus limon and Dendrobium nobile are misassembled, while Coffea arabica is not a feasible instance. Concerning the first two instances, further post-optimal analysis revealed two causes: unique regions are misassembled — this fact by itself is not a surprise since NIFS problem exclusively focuses on inverted repeat building — and/or some reads from the plant have not been filtered and were wrongly assembled into remaining contigs. On the other hand the Coffea arabica instance is unfeasible because some links are missing and/or multiplicities are under-estimated.

Quast's assessment indicated that Digitalis lanata solution is covering only 83.15% of the reference genome. However, further analysis highlighted that Quast was not able to map the contig corresponding to inverted repeat sequence to only one of the two reference genome's inverted repeats.

4 Conclusion & Discussion

In this study we show that the knowledge of the genomic regions structure in the case of chloroplasts is sufficient for the scaffolding assembly. We achieve this task by designing and implementing a linear programming tool that focuses on inverted repeats building in a circular genome context.

Although the problem is NP-Complete, our numerical experiments show that Gurobi solves real data-sets extremely fast. Our first results evaluated by Quast are very encouraging even if they reveal some issues: the first assembly step is a highly heuristic-based approach and can suffer from unremoved plant's reads that remain in contigs set. Also, estimating the multiplicities remains a hard task.

Since chloroplast genomes are multimeric (regions between inverted repeats can be considered in their both orientations) this raises the question of their suitable representation. To answer this question our strategy here is to explicitly reveal this behaviour using an enriched assembly graph where each region is one node and its associated sequence can participate in the final genome independently of the other nodes.

Acknowledgements

The authors are thankful to Abdelkader Ainouche for numerous discussions and for providing real plants and chloroplasts data.

References

[1] Zheng Xiao-Ming, Wang Junrui, Feng Li, Liu Sha, Pang Hongbo, Qi Lan, Li Jing, Sun Yan, Qiao Weihua, Zhang Lifang, Cheng Yunlian, and Yang Qingwen. Inferring the evolutionary mechanism of the chloroplast genome size by comparing whole-chloroplast genome sequences in seed plants. *Scientific Reports*, 7(1):1555, May 2017. Number: 1 Publisher: Nature Publishing Group.

[2] Natasha de Vere, Tim C. G. Rich, Sarah A. Trinder, and Charlotte Long. DNA Barcoding for Plants. In Jacqueline Batley, editor, *Plant Genotyping: Methods and Protocols*, Methods in Molecular Biology, pages 101–118. Springer, New York, NY, 2015.

[3] Jenny Tschörtner, Bin Lai, and Jens O. Krömer. Biophotovoltaics: Green Power Generation From Sunlight and Water. *Frontiers in Microbiology*, 10, 2019.

[4] Nicolas Dierckxsens, Patrick Mardulyn, and Guillaume Smits. NOVOPlasty: de novo assembly of organelle genomes from whole genome data. *Nucleic Acids Research*, 45(4):e18, February 2017.

[5] Jian-Jun Jin, Wen-Bin Yu, Jun-Bo Yang, Yu Song, Claude W. dePamphilis, Ting-Shuang Yi, and De-Zhu Li. GetOrganelle: a fast and versatile toolkit for accurate de novo assembly of organelle genomes. *Genome Biology*, 21(1):241, September 2020.

[6] Jan A. Freudenthal, Simon Pfaff, Niklas Terhoeven, Arthur Korte, Markus J. Ankenbrand, and Frank Förster. A systematic comparison of chloroplast genome assembly tools. *Genome Biology*, 21:254, September 2020.

[7] Ralph Bock and Volker Knoop, editors. *Genomics of Chloroplasts and Mitochondria*, volume 35 of *Advances in Photosynthesis and Respiration*. Springer Netherlands, Dordrecht, 2012.

[8] Victor Epain, Dominique Lavenier, and Rumen Andonov. Inverted Repeats Scaffolding for a Dedicated Chloroplast Genome Assembler, June 2022.

[9] Dan Gusfield. The RNA-Folding Problem. In Dan Gusfield, editor, *Integer Linear Programming in Computational and Systems Biology: An Entry-Level Text and Course*, pages 105–121. Cambridge University Press, Cambridge, 2019.

[10] Rayan Chikhi and Guillaume Rizk. Space-Efficient and Exact de Bruijn Graph Representation Based on a Bloom Filter. In Ben Raphael and Jijun Tang, editors, *Algorithms in Bioinformatics*, Lecture Notes in Computer Science, pages 236–248, Berlin, Heidelberg, 2012. Springer.

[11] S. F. Altschul, W. Gish, W. Miller, E. W. Myers, and D. J. Lipman. Basic local alignment search tool. *Journal of Molecular Biology*, 215(3):403–410, October 1990.

[12] Guillaume Rizk, Anaïs Gouin, Rayan Chikhi, and Claire Lemaitre. MindTheGap: integrated detection and assembly of short and long insertions. *Bioinformatics*, 30(24):3451–3457, December 2014.

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

[14] Xing-Wang Deng, Rod A. Wing, and Wilhelm Gruissem. The chloroplast genome exists in multimeric forms. *Proceedings of the National Academy of Sciences*, 86(11):4156–4160, June 1989. Publisher: National Academy of Sciences Section: Biological Sciences: Genetics.

findere: fast and precise approximate membership query

Lucas Robidou[1] and Pierre Peterlongo[1]
[1] IRISA, 263 Av. Général Leclerc, 35000, Rennes, France

Corresponding author: lucas.robidou@inria.fr

Reference paper: **Lucas Robidou, Pierre Peterlongo (2021) `findere` : fast and precise approximate membership query. *String Processing and Information Retrieval.*** http://dx.doi.org/10.1007/978-3-030-86692-1_13

1 Introduction

The indexation of the vast amount of raw sequence data available today has recently received a lot of methodological attention [1]. However, the problem remains open, and no method may presently index the hundreds of petabytes of data stored at EBI, doubling every 26 months [2].

Indexation of large volumes of genomic data generally uses k-mers (words of length k) stored in Approximate Membership Query data-structures (AMQs). However, AMQs suffer from non-avoidable false-positive calls that bias downstream analyses of methods using them. In this work we propose a simple strategy (along with its implementation) called `findere`, for reducing the false-positive rate of any AMQ indexing k-mers.

`findere` speeds up queries by a factor two and decreases the false-positive rate (or, alternatively, the size of the AMQ) by two order of magnitudes. This achievement is done without modifying the original data-structure, without generating false-negative calls and with no memory overhead.

2 Method

Given two integer values k and s, with $k > s > 0$, and a bank dataset B, the `findere` strategy consists in indexing all the s-mers from B using an AMQ. A queried k-mer is considered as "present" by `findere` if all its s-mer are found in the AMQ.

Detecting k-mers based on their s-mer leads to the apparition of a novel kind of false-positives, that we call "construction false-positives" (cFP); those cFP are negligeable with regard to the reduction of the false-positive rate achieved by `findere`.

3 `findere` implementation

We propose an implementation of `findere`, available at https://github.com/lrobidou/findere. This implementation uses a Bloom filter as its inner AMQ. However, any other AMQ implementation can be used through a simple wrapper (provided with the `findere` implementation). Although `findere` can index and query any alphabet, its implementation proposes a specialisation for genomic sequences. In this genomic context, an option to index and query canonical k-mers is also provided.

4 Results

We propose results on real biological data and on natural texts to show the practical advantages offered by `findere`, both in terms of query precision, index size, and query time.

For instance, a Bloom filter that indexes 31-mers from a metagenomic dataset with 5% False Positive Rate (FPR) requires 2.6 billion bits. With `findere`, using the same space, the FPR decreases to 0.056% (including 0.0002% cFP rate), and queries are ≈2.5 times faster.

Acknowledgements This work used HPC resources from the GenOuest bioinformatics core facility (https://www.genouest.org). The work was funded by ANR SeqDigger (ANR-19-CE45-0008).

References

[1] Camille Marchet, Christina Boucher, Simon J. Puglisi, Paul Medvedev, Mikaël Salson, and Rayan Chikhi. Data structures based on k-mers for querying large collections of sequencing data sets. *Genome Research*, 31(1):1–12, jan 2021.

[2] ENA consortium. https://www.ebi.ac.uk/ena/about/statistics, 2021.

Session 2, Amphi A

Functional and integrative genomics I

Gene Expression prediction using Deep Learning

Camille Kergal[1], Marie-Dominique Galibert[1], Catherine André[1], DoGA Consortium[¶], Christophe Hitte[1] and Thomas Derrien[1]

[1] Univ Rennes, CNRS, IGDR - UMR 6290, F-35000 Rennes, France

[¶] Members of the DoGA Consortium are listed in the Acknowledgements section of the manuscript

Corresponding Author: camille.kergal@univ-rennes1.fr

Abstract

One fundamental question in biology consists in predicting gene expression based on DNA sequence alone. To this aim, deep artificial neural networks have been recently shown to be powerful methods to predict the regulatory activity of a nucleic acid sequence and, in fine*, to assess the impact of regulatory mutations on gene expression. Yet for comparative genomic/transcriptomic studies, it is not clear whether predictive sequence models of gene expression in one species could be easily generalized to other species. The tool Basenji proposes a deep-learning approach using Convolutional Neural Networks (CNN) to predict human gene expression. We used and adapted Basenji to train a dog-specific model of gene expression using a comprehensive set of canine CAGE data (Cap Analysis of Gene Expression) produced by the DoGA consortium (n=116 experiments) corresponding to 37 core tissues. We first showed that the dog model reached similar performance than in humans with high correlations between true expression levels and predicted ones in all samples included in our model (Pearson correlations median =0.66 [min=0.34; max=0.75]). Next, we selected a subset of matched human/dog tissues, to compare the expression predictions of >1,300 orthologous dog cancer genes based on either the dog (within-species predictions) or human (cross-species predictions) models. We show that the within-species model led to higher prediction performance than the cross-species model (Pearson r = 0.65, r= 0.41, respectively). We then evaluated several genomic sequence features that could be associated with the model's effectiveness. We showed that GC content and TE content correlate significantly with the decrease in performance. Finally, given the interest of the dog as a biomedical model for cancer studies, we determined the promoter regions of 1,300 human-dog orthologous cancer genes, from which we will leverage the power of our approach to predict mutations impacting gene expression. Our model is available through github (https://github.com/ckergal/BLIMP) and will be usable via a user-friendly galaxy instance.*

Keywords Deep Learning, gene expression, regulatory variant, comparative oncology, dog

1. Introduction

Deep learning (DL) algorithms have recently attracted a lot of attention in genomics and transcriptomics since they promise to extract biological knowledge from large datasets generated by high throughput sequencing technologies in a data-driven manner [1–3]. Convolutional Neural Networks (CNN) and Recurrent Neural Network (RNN) DL approaches are particularly successful in modeling regulatory elements, outperforming traditional machine learning methods [4]. Well established DL-based tools are now able to predict the amount of RNA produced in a particular cell or tissue by learning DNA motifs such as Transcription Factor Binding Sites (TFBS) regulating gene expression levels [1]. In human genomics, the Basenji tool suite uses a CNN-based approach to learn to predict thousands of epigenetic and transcriptomic profiles only based on the human genome sequence as input [5]. Given that these computational models are fine-tuned to predict the regulatory activity of a specific DNA sequence in any tissues used as learning, they offer the benefit to evaluate the impact of non-coding regulatory variants on the level of gene expression in a tissue-specific manner, and more particularly for variants localized in essential DNA motifs learnt by the tools. This has many implications given that most known genomic loci, identified by genome wide-association studies (GWAS) to be associated with a specific trait or disease, are localized outside of protein-coding genes [6] *i.e* in regulatory regions such as promoters and/or enhancers. Using CNN-based approaches thus allows to annotate the impact of non-coding variations that may for instance create novel or alter existing TFBSs and modify gene expression.

It is yet debated whether neural networks used to train models in one species can be easily generalized to other species or whether species-specific neural networks, specifically tuned with hyperparameters (HP) optimization, would provide better predictive power. For instance, Chen *et al.* trained CNN classifiers to identify enhancer sequences based on ChIP-seq data in both human and mouse genomes [4]. Although the authors concluded that their binary classifiers generalized relatively well across species, they recognized that the features learnt by the CNN were difficult to interpret. Conversely, Cochran *et al.* recently showed that cross-species DL-based models (between human and mouse data) for TFBSs identification consistently display lower performance than within-species models [7]. For gene expression predictions, the author of Basenji tried to jointly use the mouse and human genomes to train a multi-task convolutional neural network in order to predict RNA abundance from 6,959 CAGE data (Cap Analysis of Gene Expression) from both species [8]. While this strategy led to higher performance than single training, it required to *a priori* define orthologous sequences of 131 kb between human and mouse genome that can be aligned which only represent ~40-45% of both genomes. In addition, while mouse expression data are abundant and may include specific conditions which are difficult or unethical to obtain in humans, they may not recapitulate gene expression in natural conditions.

As part of our Canine Genetics Team work in comparative oncology between human and dog, we used the Basenji framework to train and optimize a canine-specific model of gene expression, using the most comprehensive set of canine CAGE data produced by the Dog Genome Annotation Project (DoGA) consortium (www.doggenomeannotation.org). We showed that the canine-specific model has comparable performance to the human Basenji model. We then performed gene expression predictions of a canine cancer gene panel and showed that we achieved consistently higher performance when predicted with the canine model (within-species predictions) rather than using the human model (cross-species predictions). The ultimate goal of our work is to use the canine model to predict the impact of non-coding genome variations on gene expression and thus, prioritize regulatory variants associated with diseases and phenotypical traits in the dog species.

2. Materials and Methods

2.1. DoGA consortium CAGE data

In total, we used 116 samples of dog CAGE profiles (Cap Analysis of Gene Expression), which are used by the Basenji algorithm to model transcription start sites (TSS). CAGE sequencing technology allows to quantify the amount of RNA in a given biological sample at a given state and thus can explain promoter usage due to its ability to map capped 5'-ends of transcripts (TSS) with high accuracy [9]. Those 116 canine expression profiles represent 37 distinct tissues and were collected through a collaboration with the Dog Genome Annotation Project (DoGA) consortium. Sequencing of the sample was established according to the FANTOM protocol with a median number of 7.6 millions reads per sample.

2.2. Human dataset description

Human gene expression prediction model was established by Kelley et al. [5]. Training was done on a total set of 5,313 quantitative sequencing assays performed on human samples. Among those samples, 674 DNase-seq and 10 ATAC-seq experiments mapped DNA accessibility, 3,991 CHIP-seq profiled transcription factor binding or histone modifications and 638 corresponded to CAGE expression profiles. In our work, we focused on test predictions from the human model of CAGE samples only, in order to compare with similar sequencing technology as in the dog prediction model. Data processing was described in *Kelley 2020* [8]. In our comparative study between human and dog, we focused on a subset of samples of 19 human CAGE expression profiles matching dog tissues from our dog prediction model. These samples were downloaded from the FANTOM database and are available at:https://fantom.gsc.riken.jp/5/datafiles/reprocessed/hg38_latest/basic/human.tissue.hCAGE/.

2.3. Reference genome assemblies

For human data analyses, the hg38 (GRCh38) genome assembly version was used as the reference assembly. For the dog, the recent canFam4 genome assembly was used as the reference genome for mapping CAGE-seq samples and for all data processing steps of the Basenji algorithm.

2.4. Dog CAGE data Processing

In order to train and use a prediction model with the Basenji tool, data processing is necessary. For the 116 dog CAGE expression data, all BAM files were processed with the *bam_cov.py* script from the Basenji tool suite in order to transform the alignments into normalized BigWig coverage tracks, as described in *Kelley et al.* [8]. The global architecture of Basenji consists in two stages of several convolutional layers with the first stage being composed of seven blocks of convolutional layers framed with batch normalization, GELU activation function and max pooling aggregation function aiming at capturing relevant DNA motifs from the input sequence in each CAGE sample. The second stage is composed of 11 blocks of dilated convolutional layers in order to spread information across the sequence and therefore, to model long range interactions. In our application, we extracted 17,400 non-overlapping sequences of 131,072 bp (=2^{17}) from the canFam4 reference genome and further randomly distributed these sequences into three sets: train, valid and test. For valid and test sets, we assigned 10% of sequences to each and 80% for the train set as done in the Human and Mouse model [8]. Each 131 kb sequence is then aggregated into 1,024 * 116 coefficients, each one representing the coverage summed of non-overlapping 128 bp windows in the 116 CAGE expression data. Once the canine model has been trained with the 116 samples, predictions made by the model are evaluated with the remaining 10% of genome sequences (test set) by comparing them with the experimentally measured expression levels. Then, to assess the performance of the model, Pearson correlations are computed between both values (experimental and predicted).

2.5. Cancer gene panel

In order to evaluate the performance and the usability of the DL models in a comparative oncology strategy between human and dog, we established a panel of genes described to be involved in cancer and evolutionary conserved between both species. We first used a set of 1,039 human cancer genes proposed by the OncoKB database (May 2019 update) [10] and kept genes with high-confident canine orthologs according to the Ensembl Compara database (v98), as described in [11]. Then, we also included 672 genes extracted from the pan-cancer gene panel of the Nanostring company, representing orthologous genes between human and dog involved in the immune response to immuno-oncology therapies. Taking the union of these two complementary resources, a final set of 1,317 genes known to be involved in cancers and orthologous between human and dog were used to assess the performance of the model and to predict the impact of regulatory mutations in their promoter sequences (here defined as the 1,024 bp window in 5' of the Transcription Start Site - TSS). From all promoter sequences of these 1,317 genes, several features were computed such as the GC percentage and the content in transposable elements (TE). For the latter, the genomic positions of SINEs (Short Interspersed Nuclear Elements) as defined by RepeatMasker were extracted from UCSC [12] and intersected with the genomic coordinates of the cancer genes promoters using bedtools intersect version 2.25 [13]. Evolutionary conservation was analyzed by sequence comparison of the 1,024 bp promoter sequences using BLAT v35. BLAT is widely used in comparative genomics optimized for pairwise DNA-sequence alignment and was set as with parameters -minScore=10 -minIdentity=50. BLAT scores, calculated according to aligned length and sequence similarity were used to evaluate sequence alignments.

3. Results

3.1. Predicting canine gene expression using CNN

We used the Basenji framework with default hyperparameters to predict 116 CAGE (Cap Analysis of Gene Expression) data using the reference dog genome (canFam4) as input (see Methods). For each CAGE sample included in the model, we computed the Pearson correlation coefficients between the predicted expression level of the test set sequences and the experimentally measured ones from CAGE data. Using this strategy, the model (thereafter called the "DOG NO OPT" model") achieved a median Pearson correlation across tissues of 0.64 (min=0.30 for pancreas, max=0.74 for neurohypophysis) slightly lower than the human model (median R = 0.66, range=0.18 to 0.76). In order to improve the robustness of the dog model, we optimized two hyperparameters (HP) e.g. L2_scale (as a normalizing HP) and learning rate, as it has been shown that these two HP are to be optimized primarily for smaller datasets. With a grid search strategy, we produced 42 models with the same input data but with different combinations of the two HP values (**Fig 1.A**).

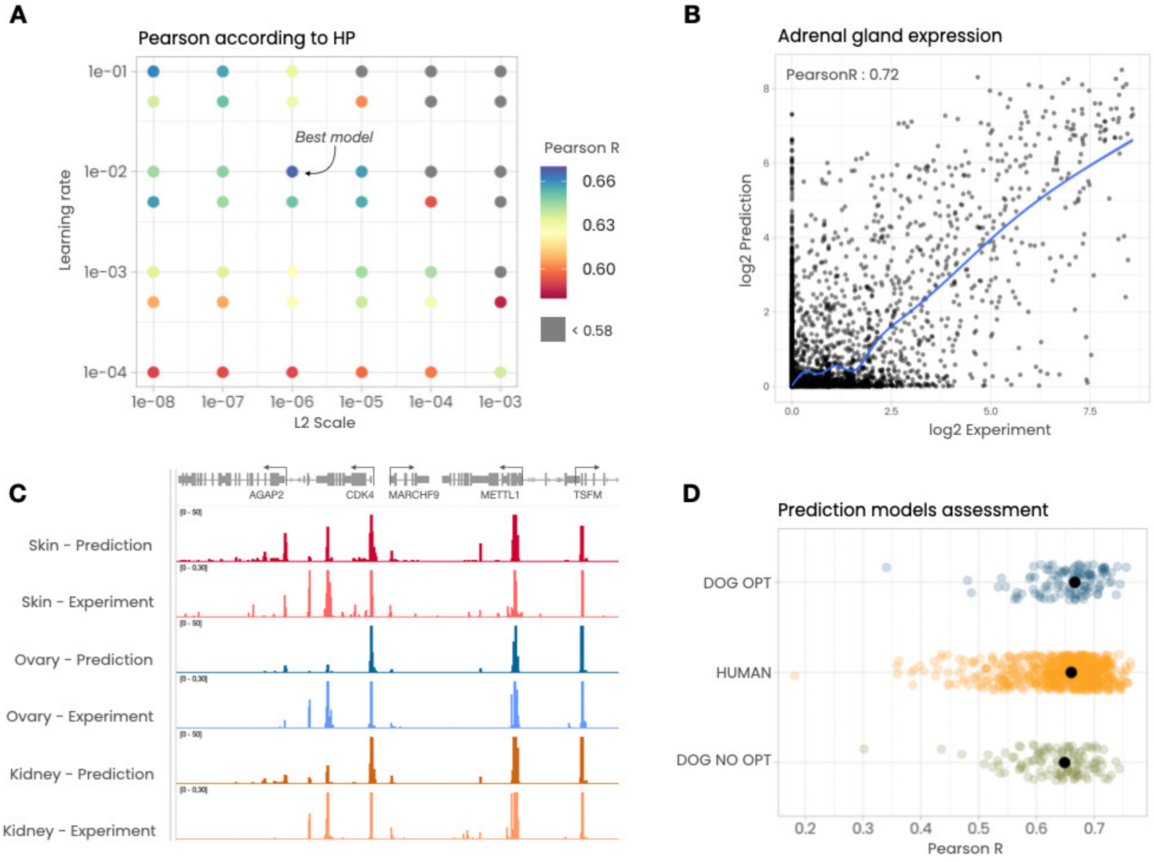


Fig 1. Canine model assessment A) Performance of the canine model as measured by Pearson correlations with respect to different values of two hyperparameters (Learning rate and L2 Scale). **B)**

Scatter plot of real/experimental (x-axis) versus predicted (y-axis) log2 sequence expressions in the adrenal gland CAGE sample. The blue line represents the Generalized Additive Model (GAM) regression between experiment and prediction values. **C)** IGV (Integrative Genomics Viewer) [14] representation of a dog canFam4 genomic region encompassing gene track (top layer) and from top to bottom predicted (dark) and experimental (light) CAGE tracks of skin (red), ovary (blue) and kidney (orange) samples. **D)** Pearson coefficients in all tissues for non-optimized dog model (DOG NO OPT), human (HUMAN) and dog model after optimization (DOG OPT). Median Pearson R are represented as black dots.

All resulting models were then ranked according to their median Pearson coefficients and the model whose HP set led to the more accurate predictions was selected. For each sample, we could derive correlations between true and predicted expression values of the tested sequences as illustrated for the adrenal gland corresponding sample in **Fig 1.B**. We also used a complementary strategy to visualize the quality of the model by generating genome-wide tracks of predicted canine CAGE profiles and comparing them with experimentally derived CAGE profiles (**Fig 1.C**). Combining all samples, the final model (called "Optimized model or DOG OPT") achieved a median correlation of 0.66 (min=0.34, max=0.75) (**Fig 1.D**).

We also compared our canine model performance with the human prediction model from the original Basenji tool described by *Kelley D.R. 2020* [8]. In order not to bias the comparison with different sequencing technologies, only values predicted by the human model from CAGE samples were used for the comparison with the canine prediction of CAGE gene expression (**Fig 1.D**). Altogether, the dog optimized led to slightly higher performance than the human model.

3.2. Comparison of within-species versus cross-species models of gene expression

At the transcriptional level, human and dog protein-coding gene expression profiles are globally conserved. It has been shown that orthologous gene expression profiles cluster preferentially by tissues rather than by species [15, 16]. Comparative transcriptomic studies motivated us to investigate whether learned DNA motifs predictive of gene expression in one species could also be used to predict gene expression in another species. In other words, given a test set of canine genes, we computed the performance of the dog model (within-species predictions) with respect to the human model (cross-species predictions) in order to predict their expression levels in matched tissues between both species. We used 1,317 canine genes conserved in human (see Methods) and for each similar tissue, we measured the performance of both the dog and the human model to predict their expression levels **(Fig 2.A)**.

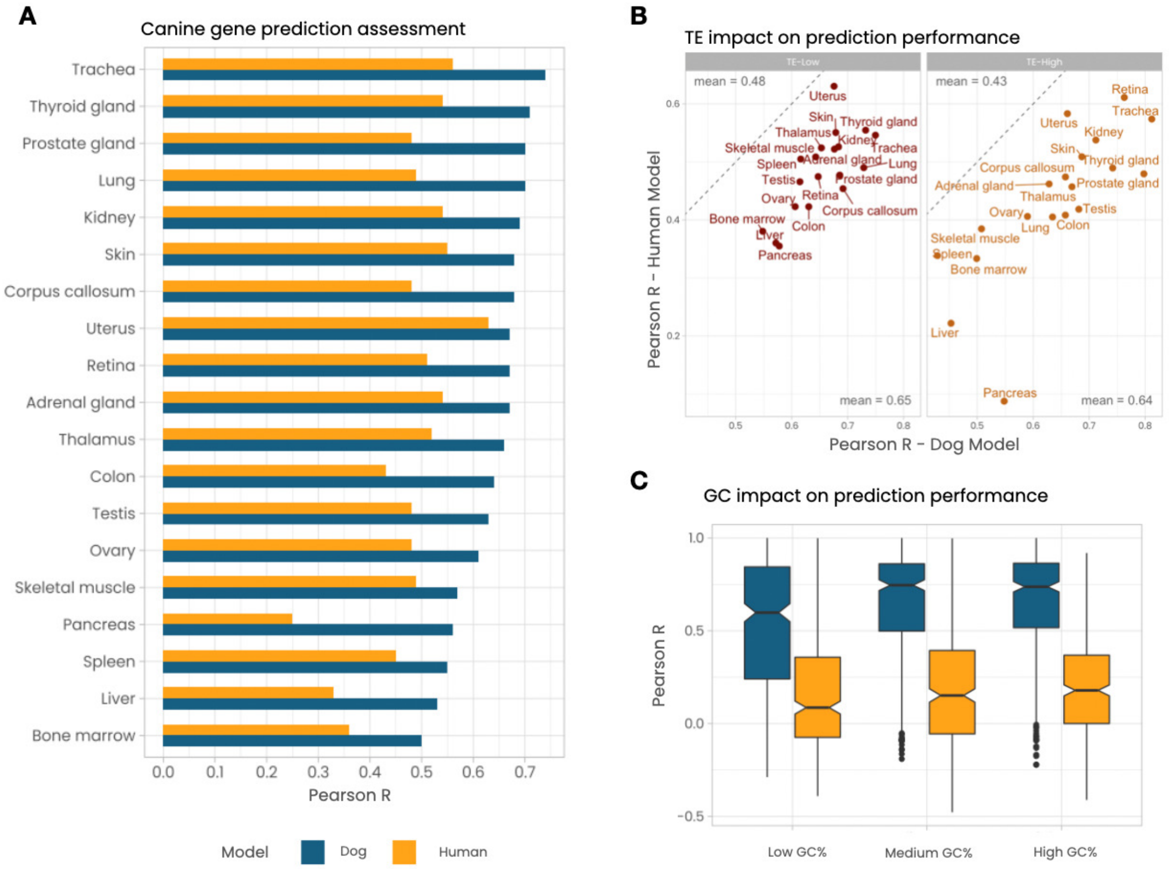


Fig 2. Within-species versus cross-species prediction. A) Performance of the DOG model (within-species) (blue) versus the HUMAN model (cross-species) (yellow) for gene expression of canine cancer genes across 19 matched dog/human tissues. For each tissue, we computed the Hotelling-Williams test to assess the significance of the difference between the model predictions. Each p-value is highly significant. **B)** Influence of gene promoter contents in Transposable Element (TE) (low = left panel; high = right panel) for within-species prediction (x-axis) versus cross-species prediction (y-axis); dotted diagonal represents the line where y=x . **C)** Influence on Pearson correlation of gene promoter content in GC (low, medium and high according to ⅓ and ⅔ percentiles GC%) for within-species prediction (blue boxes) and cross-species prediction (yellow boxes).

This study highlighted that in all of the 19 tissues considered, the dog model always exhibited higher Pearson coefficients than those obtained with the human model (median R_{within} = 0.67 versus R_{cross} = 0.49). Next, we investigated the genomic features that could explain the lower predictive power of the human model. For the 1,317 canine promoter sequences, we computed several features such as GC content, evolutionary conservation and content in transposable elements (TEs) (see Methods). For the latter, we categorized genes as whether they contain canine SINE elements or not in their promoter sequences and computed Pearson coefficients separately for these two sets of genes for the two models. The performance of the dog model was not significantly different for genes having or not TE in their promoters (R=0.65 versus R=0.64 respectively) **(Fig 2.B)**. This could be expected given that the dog model was trained on the entire dog genome sequence which included TE elements. Conversely, the human model performed poorly for dog genes containing TE in their promoters as compared to genes without TEs (R=0.43 versus R=0.48 respectively), probably highlighting canine-specific SINEs important for gene expression regulation but not learnt by the human model. To assess the impact of GC content of gene promoter on prediction performance, gene promoter sequences were split in three categories (low, medium and high GC) with respect to ⅓ and ⅔ percentiles GC **(Fig 2.C)**. With both prediction models, performance was improved for promoters exhibiting high and medium GC content although the difference was more pronounced with the within-species model compared to the cross-species model (R_{cross} = 0.12, R_{cross} = 0.15 and R_{cross} = 0.18, R_{within}= 0.57, R_{cross} = 0.74 and R_{cross} = 0.73 for low, medium and high GC content, respectively). Interestingly in dogs, it is known that gene promoters are highly enriched in GC content, in part due to the loss of the *PRDM9* gene [17]. Finally, we also compared the influence of sequence conservation in the performance of the cross-species model. As for GC content, gene promoters were separated as lowly and highly conserved with respect to median blat score (see Methods). As expected, performance of the cross-species model was found lower for genes with lower level of sequence conservation than those with a higher level of sequence conservation (R_{cross}=0.34 and 0.45, respectively).

4. Conclusion - Discussion

We have developed a predictive model that can process a genome sequence to better understand gene expression in dogs. We present a CNN-based canine-specific model of gene expression using the most comprehensive set of CAGE data provided by the DoGA consortium. We showed that this dog model has better performance for predicting dog gene expression compared to a cross-species strategy using a human trained model. Hence, we provide a canine-specific model that can be used to predict gene expression and their variations for the largest collection of tissue (n=37) currently available. This computational tool is designed to be used by the scientific community that relies on the dog genetic model to decipher complex diseases and phenotypes mapping. Although the cross-species model displays lower performance, under specific conditions (GC and TE content, sequence conservation), a sub-optimal use of the cross-species model can be envisaged to extend its application to the several hundreds of tissues and cell lines included in the human model.

While this CNN-based model reached good correlations between predicted and experimentally measured expression levels, recent deep learning architecture involving transformers have been shown to outperform CNN particularly for capturing longer interaction between genes and their regulatory elements [18]. It could thus be interesting to apply such approaches for dog gene expression prediction.

The ultimate goal of gene expression models consists in learning the DNA regulatory code responsible for gene expression in a tissue-specific manner and thus to categorize the complexity of non-coding mutations and prioritize their impacts and roles. In our work, we have defined and extracted the gene promoter sequence of 1,317 genes involved in cancers. Using an *in silico* saturation mutagenesis analysis [19] of the promoter sequences, we seek to identify the regulatory variations between humans and dogs that will be predicted by both models that *a priori* significantly alter gene expression levels in both species. This aspect will be covered in an extended version of the work.

Acknowledgements

Authors thank the bioinformatics core facility Genouest (https://www.genouest.org/) and the Dog Genome Annotation (DoGA) Consortium (César L. Araujo, Milla Salonen, Riika Sarviaho, Julia Niskanen, Sruthi Hundi, Jenni Puurunen, Sini Sulkama, Sini Karjalainen, Antti Sukura, Pernilla Syrjä, Niina Airas, Henna Pekkarinen, Ilona Kareinen, Anna Knuuttila, Heli Nordgren, Karoliina Hagner, Tarja Pääkkönen, Kaarel Krjutskov, Sini Ezer, Shintaro Katayama, Masahito Yoshihara, Auli Saarinen, Matthias Hörtenhuber, Amitha Raman, Irene Stevens, Maria Kaukonen, Ileana B. Quintero, Abdul Kadir Mukarram, Marjo K. Hytönen, Kaisa Kyöstilä, Meharji Arumili, Carsten O. Daub, Juha Kere, Hannes Lohi) for the CAGE data. David R. Kelley for helpful suggestion during elaboration of the dog prediction model (http://www.davidrkelley.com/)

References

[1] J. Zou, M. Huss, A. Abid, P. Mohammadi, A. Torkamani, and A. Telenti, “A primer on deep learning in genomics,” *Nature Genetics*, 2018, doi: 10.1038/s41588-018-0295-5.

[2] J. Zhou, C. L. Theesfeld, K. Yao, K. M. Chen, A. K. Wong, and O. G. Troyanskaya, “Deep learning sequence-based ab initio prediction of variant effects on expression and disease risk,” *Nat Genet*, vol. 50, no. 8, pp. 1171–1179, Aug. 2018, doi: 10.1038/s41588-018-0160-6.

[3] G. Eraslan, Ž. Avsec, J. Gagneur, and F. J. Theis, “Deep learning: new computational modelling techniques for genomics,” *Nat Rev Genet*, vol. 20, no. 7, pp. 389–403, Jul. 2019, doi: 10.1038/s41576-019-0122-6.

[4] L. Chen, A. E. Fish, and J. A. Capra, “Prediction of gene regulatory enhancers across species reveals evolutionarily conserved sequence properties,” *PLoS Comput Biol*, vol. 14, no. 10, p. e1006484, Oct. 2018, doi: 10.1371/journal.pcbi.1006484.

[5] D. R. Kelley, Y. A. Reshef, M. Bileschi, D. Belanger, C. Y. McLean, and J. Snoek, “Sequential regulatory activity prediction across chromosomes with convolutional neural networks.,” *Genome research*, vol. 28, no. 5, pp. 739–750, 2018, doi: 10.1101/gr.227819.117.

[6] L. A. Hindorff *et al.*, “Potential etiologic and functional implications of genome-wide association loci for human diseases and traits,” *Proceedings of the National Academy of Sciences*, vol. 106, no. 23, pp. 9362–9367, 2009, doi: 10.1073/pnas.0903103106.

[7] K. Cochran *et al.*, “Domain-adaptive neural networks improve cross-species prediction of transcription factor binding,” *Genome Res.*, vol. 32, no. 3, pp. 512–523, Mar. 2022, doi: 10.1101/gr.275394.121.

[8] D. R. Kelley, “Cross-species regulatory sequence activity prediction,” *PLOS Computational Biology*, vol. 16, no. 7, p. e1008050, Jul. 2020, doi: 10.1371/journal.pcbi.1008050.

[9] M. Lizio *et al.*, “Monitoring transcription initiation activities in rat and dog,” *Sci Data*, vol. 4, no. 1, p. 170173, Nov. 2017, doi: 10.1038/sdata.2017.173.

[10] D. Chakravarty *et al.*, “OncoKB: A Precision Oncology Knowledge Base,” *JCO Precis Oncol*, vol. 2017, Jul. 2017, doi: 10.1200/PO.17.00011.

[11] J. Herrero *et al.*, “Ensembl comparative genomics resources,” *Database : the journal of biological databases and curation*, vol. 2016, pp. 1–17, 2016, doi: 10.1093/database/bav096.

[12] J. Navarro Gonzalez *et al.*, “The UCSC Genome Browser database: 2021 update,” *Nucleic Acids Research*, vol. 49, no. D1, pp. D1046–D1057, Jan. 2021, doi: 10.1093/nar/gkaa1070.

[13] A. R. Quinlan and I. M. Hall, “BEDTools: a flexible suite of utilities for comparing genomic features,” *BIOINFORMATICS APPLICATIONS NOTE*, vol. 26, no. 6, pp. 841–842, 2010, doi: 10.1093/bioinformatics/btq033.

[14] J. T. Robinson, H. Thorvaldsdóttir, W. Winckler, M. Guttman, E. S. Lander, G. Getz, J. P. Mesirov. Integrative Genomics Viewer. Nature Biotechnology 29, 24–26 (2011).

[15] C. Le Béguec *et al.*, “Characterisation and functional predictions of canine long non-coding RNAs.,” *Scientific reports*, vol. 8, no. 1, p. 13444, Sep. 2018, doi: 10.1038/s41598-018-31770-2.

[16] A. Breschi, T. R. Gingeras, and R. Guigó, “Comparative transcriptomics in human and mouse,” *Nature Reviews Genetics*, vol. 18, no. 7, pp. 425–440, 2017, doi: 10.1038/nrg.2017.19.

[17] C. Wang *et al.*, “A novel canine reference genome resolves genomic architecture and uncovers transcript complexity,” *Communications Biology*, vol. 4, no. 1, p. 185, Dec. 2021, doi: 10.1038/s42003-021-01698-x.

[18] Ž. Avsec *et al.*, “Effective gene expression prediction from sequence by integrating long-range interactions,” *Nat Methods*, pp. 1–8, Oct. 2021, doi: 10.1038/s41592-021-01252-x.

[19] S. Nair, A. Shrikumar, J. Schreiber, and A. Kundaje, “fastISM: Performant in-silico saturation mutagenesis for convolutional neural networks,” *Bioinformatics*, p. btac135, Mar. 2022, doi: 10.1093/bioinformatics/btac135.

Session 2, Amphi B

Knowledge representation, databases & visualization I

FORUM: Knowledge Graph (KG) for semantic representation and inference of relations between chemicals and biomedical concepts

Maxime Delmas[1], Olivier Filangi[2], Nils Paulhe[3], Florence Vinson[1], Christophe Duperier[3], William Garrier[4], P-E Saunier[4], Yoann Pitarch[5], Fabien Jourdan[1], Franck Giacomoni[3] and Clément Frainay[1]

[1] Toxalim (Research Center in Food Toxicology), Université de Toulouse, INRAE, ENVT, INP-Purpan, UPS, 31300 Toulouse, France

[2] IGEPP, INRAE, Institut Agro, Université de Rennes, Domaine de la Motte, 35653 Le Rheu, France

[3] Université Clermont Auvergne, INRAE, UNH, Plateforme d'Exploration du Métabolisme, MetaboHUB Clermont, F-63000 Clermont-Ferrand, France

[4] ISIMA, Campus des Cézeaux, 63177 Aubière, France

[5] IRIT, Université de Toulouse, Cours Rose Dieng-Kuntz, 31400 Toulouse, France

Corresponding author: maxime.delmas@inrae.fr

Reference paper: **Delmas *et al.* (2021) FORUM: building a Knowledge Graph from public databases and scientific literature to extract associations between chemicals and diseases.** https://doi.org/10.1093/bioinformatics/btab627

Metabolomics applied to human health allows identifying lists of small molecules (metabolic profiles) whose concentrations are significantly modified. Metabolic profiles are of high importance in human phenotyping studies, for instance to stratify patients or identify biomarkers. Nevertheless, one of the main challenges is to associate these lists of discriminant metabolites to potential pathological effects. This step usually requires a time-consuming dive in the scientific literature, which comprises more than 31 million of citations at the end of 2021 (PubMed repository only). Hence, to assist researchers, we introduce FORUM [1]: an open Knowledge Graph (KG) built from a large-scale federation of life science databases and scientific literature repositories, from which have been postulated relations between chemicals and biomedical concepts.

We collected co-mentions between chemical compounds and biomedical concepts in the literature and formalize them as statements in the KG, along with a semantic descriptions of these entities. This allows us to apply automated ontological reasoning and draw implicit connections through the literature between more broader concepts, representative of families of chemicals or disorders. FORUM thus provides different levels of abstraction, which may be particularly useful regarding hypothesis generation for signature interpretation. We extracted significant relations based on their co-mention frequency and instantiated them as a new layer of relations in the KG to support results interpretation and further inquiries. This was also extended to relations between biomedical concepts themselves (like therapeutic actions and diseases) to supplement the KG and enhance hypothesis discovery from literature [2]. To this end, relations are not limited to what is modeled in the FORUM KG but can be extended through several other Linked Open Data datasets using federated queries, thus substantially increasing the potential connections that can be established.

Containing more than 9 billion factual statements from which have been inferred more than 9 million relations between compounds and biomedical concepts, FORUM stands as a new FAIR and valuable resource for a deeper and comprehensive analysis of metabolomics data. We showcase the benefit of FORUM by comparing our results to previously published test-cases as well as exploring complex paths of relations to propose new hypotheses. A web interface to browse and download the extracted relations, as well as a SPARQL endpoint to directly probe FORUM KG, are available at https://forum-webapp.semantic-metabolomics.fr.

References

[1] Maxime Delmas, Olivier Filangi, Nils Paulhe, Florence Vinson, Christophe Duperier, William Garrier, Paul-Emeric Saunier, Yoann Pitarch, Fabien Jourdan, Franck Giacomoni, and Clément Frainay. FORUM: building a Knowledge Graph from public databases and scientific literature to extract associations between chemicals and diseases. *Bioinformatics*, 37(21):3896–3904, November 2021.

[2] Don R. Swanson and Neil R. Smalheiser. An interactive system for finding complementary literatures: a stimulus to scientific discovery. *Artificial Intelligence*, 91(2):183–203, April 1997. MAG ID: 2024030746.

Session 2, Amphi C

Statistics, machine learning, AI & image analysis I

TopoFun: a machine learning method to improve the functional similarity of gene co-expression modules

Ali Janbain[1,2], Christelle Reynès[1], Zainab Assaghir[2], Hassan Zeineddine[2], Robert Sabatier[1] and Laurent Journot[1,3]

[1] IGF-Institut de Génomique Fonctionnelle, Univ. Montpellier, CNRS, INSERM, 34094, Montpellier, France

[2] Applied Mathematics Department, Lebanese University, 1003, Beirut, Lebanon

[3] MGX-Montpellier GenomiX, Univ. Montpellier, CNRS, INSERM, 34094, Montpellier, France

Corresponding Author: laurent.journot@igf.cnrs.fr

Janbain *et al.* (2021) TopoFun: a machine learning method to improve the functional similarity of gene co-expression modules, NAR Genom Bioinform, 3:lqab103. https://doi.org/10.1093/nargab/lqab103

1. Introduction

Following the classical definition of functional modules (FMs) by Hartwell *et al.* [1], the methods aimed at identifying FMs in gene co-expression networks all assume that the members of the same FM are more densely connected than those across FMs. This strategy does not take into account the shared functional annotations of FM members. We decided to make no hypothesis about FMs' topology and learned their properties using machine learning on known FMs. This knowledge was then used to identify new FMs and to find new members of known FMs.

2. Methods

We started from the 978 Gene Ontology Biological Processes (GO-BPs) with 20-500 genes as prototypic FMs and generated an equivalent number of random modules (RMs). We computed 12 descriptors of the modules' topology, and used LASSO to select those that best discriminated FMs and RMs. We performed Linear Discriminant Analysis to construct a topological score ($\text{Score}_{\text{Topo}}$) that predicted the type of a module, random- or functional-like. Following Wang *et al.* [2], we designed a functional similarity score ($\text{Score}_{\text{Fun}}$) based on the distance in the GO tree of the annotations of the genes that constitute a module. $\text{Score}_{\text{Topo}}$ and $\text{Score}_{\text{Fun}}$ estimate to which extent a module is made of genes as functionally related as those of GO-BPs. We combined both scores in a fitness function that ranked candidate modules and used it in a genetic algorithm we named TopoFun to explore the neighborhood of sets of genes of interest.

3. Results

To illustrate the use of TopoFun, we started from a subset of the GO-BPs and showed that TopoFun efficiently retrieved genes that we omitted, and aggregated a number of novel genes while improving module topology and functional similarity. Using an independent protein-protein interaction database (STRING), we confirmed that the novel genes were functionally related to the original gene set. Finally, we evaluated the functional modules produced by WGCNA [3], the reference method in co-expression network algorithms. WGCNA modules displayed an exceptional topology, better than that of known GO-BPs. However, they lacked functional coherence and were not significantly different from RMs from that point of view. We used TopoFun to improve the WGCNA modules' functional similarity and produced modules with a slightly improved topology and much higher functional consistency.

References

1. Leland H. Hartwell, John J. Hopfield, Stanislas Leibler, and Andrew W. Murray. From molecular to modular cell biology. *Nature*, 402:C47–52, 1999.
2. James Z. Wang, Zhidian Du, Rapeeporn Payattakool, Philip S. Yu, and Chin-Fu Chen. A new method to measure the semantic similarity of GO terms. *Bioinformatics*, 23:1274–81, 2007.
3. Peter Langfelder, and Steve Horvath.WGCNA: an R package for weighted correlation network analysis. *BMC Bioinformatics*, 9:559, 2008.

GraphGONet: a self-explaining neural network encapsulating the Gene Ontology graph for phenotype prediction on gene expression

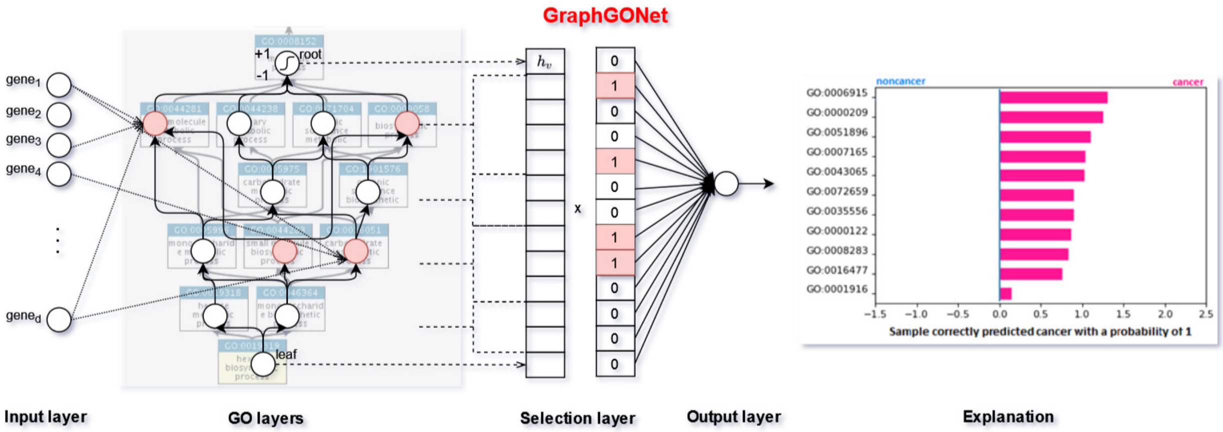
Victoria BOURGEAIS, Farida ZEHRAOUI and Blaise HANCZAR
Laboratory IBISC, Université Paris-Saclay (Univ. Évry), 36 Rue du Pelvoux, 91020, Évry-Courcouronnes, France

Corresponding author: victoria.bourgeais@universite-paris-saclay.fr

***Reference paper:* Bourgeais *et al.* (2022) GraphGONet: a self-explaining neural network encapsulating the Gene Ontology graph for phenotype prediction on gene expression. *Bioinformatics*. 10.1093/bioinformatics/btac147**

Medical care is becoming more and more specific to patients' needs due to the increased availability of omics data such as Gene Expression (GE) profiles. The application to these data of sophisticated Machine Learning (ML) models, in particular Deep Learning (DL), can improve the field of precision medicine. However, their use in clinics is limited as most of these algorithms are considered black-boxes. This means that their predictions are not accompanied by any explanations. Therefore, an important issue today is making ML algorithms interpretable. In addition, the production of accurate and intelligible predictions can benefit from the inclusion of domain knowledge; thus, knowledge-based DL models appear to be a promising solution.

In this paper, we propose GraphGONet, where the Gene Ontology (GO) is encapsulated in the hidden layers of a new self-explaining Neural Network (NN). Each neuron in the layers represents a GO term and each connection between the neurons represents a relation in GO. Our method is composed of a particular type of Feedforward Neural Network (FFNN) and a special case of Graph Neural Network (GNN). In our proposed FFNN, the input layer can be connected to each hidden layer and all hidden layers are connected to the output layer through a selection layer. The propagation of the signal through the GO layers is inspired by the propagation rules in GNN, i.e, a neuron aggregates the information respectively from its set of connected genes and its child GO terms. In addition to accurate predictions, our model can produce automatic explanations. Specifically, the last layer of the NN contains the set of the most important neurons for the prediction and their associated GO terms. This set directly explains the outcome of a patient, thus an enrichment test is no longer required.



The experiments described in the paper confirm that our model not only performs as accurately as the state-of-the-art (non-explainable ones) but also automatically produces stable and intelligible explanations composed of the biological concepts with the highest contribution to the prediction. This feature allows experts to use our tool in a medical setting. Different types and levels of explainability can be proposed depending on the end user's expectations and the reliability of the system can be estimated from the performance and explainability offered by the model. In future work, we plan to include other ontologies, such as pathways [1], that could enrich the biological explanations. The code source of GraphGONet is available online.

References

[1] Jie Hao, Youngsoon Kim, Tae-Kyung Kim, and Mingon Kang. PASNet: pathway-associated sparse deep neural network for prognosis prediction from high-throughput data. *BMC Bioinformatics*, 19(1):510, 2018.

Adaptive splines-based logistic regression with a ReLU neural network

Marie GUYOMARD[1], Susana BARBOSA[2] and Lionel FILLATRE[1]
[1] University Côte d'Azur, Laboratory I3S, France
[2] University Côte d'Azur, Laboratory IPMC, France

Corresponding author: guyomard@i3s.unice.fr

Abstract *This paper proposes a non-linear binary classification model. Although linear classification methods are very popular in the field of personalized medicine because of their interpretability, they have proven to be too restrictive. Doctors are convinced of the need to quantify threshold effects for better predictions. Nevertheless, non-linear methods that can be found in the state of the art are not able to automate the segmentation of variables (Segmented Logistic Regression) or are difficult to interpret (Random Forests or Neural Networks). We propose a Neural Network that fully realizes a non-linear logistic regression. The score function of the logistic regression, initially linear, is replaced by a piecewise linear function, modeled by spline functions. The particular architecture of this network automates the segmentation of the variables and guarantees its operational relevance as well as the explicability of its calculated predictions.*

Keywords Non-linear Classification, Splines approximation, Neural Networks.

1 Introduction

The use of Artificial Intelligence in the medical field continues to progress. Machine learning models for classification allow in many practical cases to avoid invasive methods, such as biopsies, to provide an accurate diagnosis. The use of Logistic Regression (LR) is widespread. For example, in [1] the LR is used to predict the development of non-alcoholic cirrhosis. Unlike methods such as boosting, random forests, and neural networks (NNs), the model estimated by LR is easily interpretable since it depends on a linear combination of explanatory variables.

Nevertheless, doctors believe that incorporating nonlinear phenomena in the modeling, such as threshold effects on certain variables, would increase the predictive performance. For example, a significantly high cholesterol level could be a risk factor for developing a disease while a low cholesterol level would decrease this risk. For this purpose, a promising track is the use of a LR that exploits a non-linear model based on piece-wise linear splines. These splines divide the domain of definition of the explanatory variables into several segments and, on each segment, perform a linear approximation.

The major difficulty of using an approximation with splines relies on the choice of the bounds of each segment, called knots of the spline. It has been shown in [2,3] that jointly optimizing the knots and the linear approximation associated with each segment is difficult. The fixed splines based-LR avoids these optimization issues by fixing a priori the number and the value of the knots. The choice of knots is therefore rather arbitrary and the model becomes frozen. An alternative is the MARS (Multivariate Adaptative Regression Splines) model which proposes an adaptive method to compute the knots. Those are computed recursively in order to progressively improve the performance of the model. The global performance of the method is not directly optimized but in a greedy and sub-optimal way. Recently, a third approach has emerged : it establishes a rigorous bridge between deep neural networks (NNs) and the theory of piecewise linear spline functions. Contrary to the MARS model, in NNs a global criterion is minimized to learn the segmentation. Unfortunately, the segmentation produced by an NN is very complex and therefore it becomes impossible to easily interpret the impact of explanatory variables in the prediction.

This work has been supported by the French government, through the UCA DS4H Investments in the Future project managed by the National Research Agency (ANR) with the reference number ANR-17-EURE-0004.

The main contribution of this paper consists in developing an NN inspired by the MARS model in order to combine the advantages of MARS and NNs: the minimization of a global criterion to obtain an NN with an easily explicable architecture. A second contribution consists in proposing an explicit algorithm to train this NN. Indeed, the authors in [4] demonstrate that it is possible to approximate MARS models by NNs but their approach, totally theoretical, does not propose any algorithm to train the NN. Finally, we compare our NN to the previously mentioned methods on a simulated and a real datasets. Our NN presents comparable or superior performances in prediction to the other approaches while being fully explainable.

This paper is structured as follows. The section 2 introduces the studied prediction problem. The section 3 describes our NN architecture. The section 4 presents the experimental results. Finally, the section 5 concludes the article.

2 Problem Statement

We have N independent and identically distributed pairs $(x^{(i)}, y^{(i)})$ where $x^{(i)}$ is the vector of explanatory variables and $y^{(i)} = \{0, 1\}$ is the binary label to predict. The notation (X, Y) denotes the pair of random variables from which the pairs $(x^{(i)}, y^{(i)})$ are derived. These data are used to train and test all the models implemented in the article.

2.1 Bayesian binary classification

A Bayesian Maximum a Posteriori (MAP) classifier assigns a label y to a sample $x = [x_1, \dots, x_d]$ based on the decision rule $\delta : \mathbb{R}^d \mapsto [0, 1]$ defined by $\delta(x) = \hat{\mathbb{P}}(Y = y|X = x)$ where $\hat{\mathbb{P}}(Y|X)$ is an estimate of the conditional a posteriori probability. The LR is the most widely used decision rule for this kind of problem in the medical field. It is written as

$$\delta^{\mathrm{RL}}(x) = \sigma(\psi(x)) = \frac{1}{1 + \exp\left(-\psi(x)\right)}, \tag{1}$$

where $\sigma(\cdot)$ is the logistic function and $\psi(x)$, called the score function, is a linear function $\psi(x) = \beta^\top x$ where $\beta = [\beta_1, \dots, \beta_d] \in \mathbb{R}^d$ is a vector of coefficients and $\beta^\top$ denotes the transposed vector β. Each coefficient β_i quantifies the impact of the i^{th} component of the vector x on the probability of choosing the class $y = 1$. This model is very appreciated for its simplicity and its explicability.

2.2 MARS score function of order 1

In order to obtain a non-linear but still explainable score function, a relevant modeling is brought by the MARS approach [5]. This model is based on an approximation with adaptive splines of the score function:

$$\psi^{\mathrm{MARS}}(x) = \sum_{m=1}^{M} \beta_m h_m(x), \tag{2}$$

where $h_m(x)$ is a spline function of the form

$$\begin{align} h_m(x) &= \left[s_m(x_{v(m)} - b_m)\right]_+ \tag{3} \\ &= \begin{cases} \max\{0, x_{v(m)} - b_m\}, & \text{if } s_m = 1, \\ \max\{0, b_m - x_{v(m)}\}, & \text{if } s_m = -1. \end{cases} \tag{4} \end{align}$$

The notation $[t]_+ = \max\{0, t\}$ denotes the ReLU function. The function $h_m(x)$ depends on the $v(m)^{\text{th}}$ component $x_{v(m)}$ of the vector x. The real b_m is the knot of the spline. The integer $s_m \in \{-1, 1\}$ used in conjunction with the ReLU function allows to cancel the left part or the right part of $h_m(x)$ as explained in (4).

The MARS approach learns the $h_m(x)$ functions sequentially. At each iteration $m \in \{1, \dots, M\}$, the spline function $h_m(\cdot)$ that best reduces the learning error is added. The recursive and adaptive segmentation of the MARS approach is thus similar to that of decision trees. If $v(m) \neq k$ for all m, then the k^{th} component of x will never be included in the model. The MARS approach is based on a greedy optimization algorithm whose global optimality is not established. The recursivity of the model makes the segmentation of variables uncontrollable. It is possible that a same variable is segmented a large number of times. However, we know from doctors' feedback that over-segmenting a biological variable is not relevant.

2.3 ReLU Neural Networks

The non-linear approach that is currently very widespread is based on a score function produced by a ReLU NN [6]. In this paper, we consider only single-hidden-layer NNs that are written as

$$\psi^{\mathrm{NN}}(x) = \beta_0 + \beta^\top [Wx + b]_+, \tag{5}$$

where $\beta \in \mathbb{R}^p$, $W \in \mathbb{R}^{p\times d}$ is the matrix of weights, $b \in \mathbb{R}^p$ is the vector of biases, and $[z]_+$ denotes the vector z where the function $[\cdot]_+$ has been applied to each component. The hidden layer has p neurons. According to [7], deep networks with the ReLU activation function introduce a partitioning of $\mathbb{R}^d$ that is equivalent to an approximation with multidimensional splines. However, this partitioning is very complex and generally unexplainable. The figure 2-e illustrates this partitioning when $d = 2$ with $p = 30$ neurons. The straight lines, almost always oblique, intersect and cut the space into polyhedra with very diverse geometrical shapes. This division explains the flexibility of an NN but also why an NN is considered as a "black box".

3 Neural Network NN-MARS

In order to benefit from the advantage of training a neural network (minimization of a global criterion with a gradient descent) and to keep an explicability close to the MARS model, we propose in this section a piecewise continuous score function modeled with a NN. Our model $\psi^{\text{NN-MARS}}(x)$ is written as

$$\psi^{\text{NN-MARS}}(x) = \beta_0 + \sum_{j=1}^{d} g_j(x_j), \tag{6}$$

$$g_j(t) = \beta_{j1}[b_{j1} - t]_+ + \beta_{j2}[t - b_{j2}]_+, \quad t \in \mathbb{R}. \tag{7}$$

In (6), the real nonlinear function $g_j(\cdot)$ is applied to x_j, the j^{th} component of the vector $x \in \mathbb{R}^d$. The function $g_j(t)$ corresponds to a pair of neurons working together: the first neuron is a non-zero spline before the knot value b_{j1} and the second one is a non-zero spline after the knot value b_{j2}. As a result, $g_j(t)$ models functions with a pattern composed of 3 linear segments as illustrated in the gray box on Figure 1. In this figure, the variable X_1 can represent for example the weight of a patient. Being underweight ($X_1 < b_{11}$) or overweight ($X_1 > b_{12}$) increases the probability of developing the pathology. In contrast, between these two intervals, the impact of weight on disease is negligible. The segmentation of the network is therefore controlled: at most 3 segments are created for each descriptive variable.

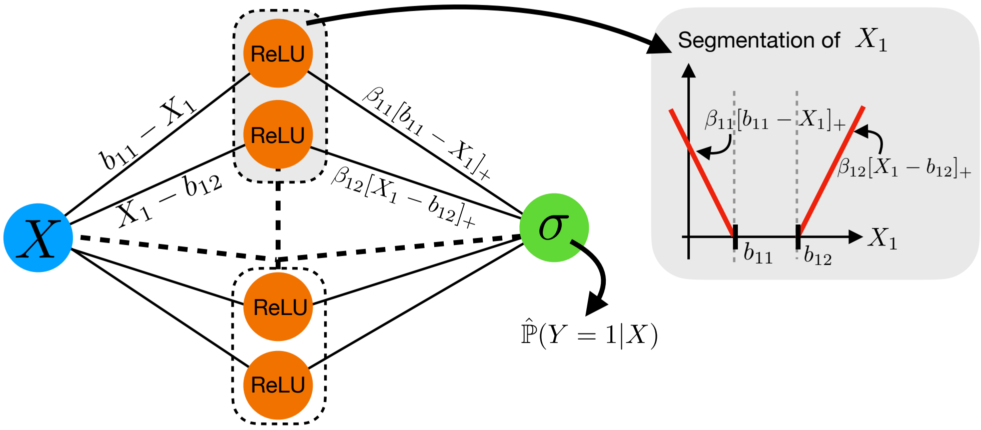

ReLU
ReLU
ReLU
ReLU
X
σ
$b_{11} - X_1$
$X_1 - b_{12}$
$\beta_{11}[b_{11} - X_1]_+$
$\beta_{12}[X_1 - b_{12}]_+$
$\hat{\mathbb{P}}(Y = 1|X)$
Segmentation of X_1
$\beta_{11}[b_{11} - X_1]_+$
$\beta_{12}[X_1 - b_{12}]_+$
X_1
b_{11}
b_{12}

Fig. 1. Architecture of NN-MARS: the inputs in blue, the hidden layer in orange and the estimated labels in green.

Furthermore, the nature of the segmentation operated by the NN-MARS is also controlled. Contrary to classical ReLU NNs which create oblique regions by linearly combining the components of x, the NN-MARS cuts the components of x independently. It relies on hyperplanes that are orthogonal to the canonical basis of the $\mathbb{R}^d$ space, just as decision trees or the MARS model do, as illustrated in Figure 2-f. The partitioning of the $\mathbb{R}^d$ space is done with hypercubes and not polyhedra with complex shapes. The decision rule obtained is easily interpretable: the score function is linear on each hypercube. In practice, this is equivalent to performing a local LR by thresholding the components of the vector x. The function $\psi^{\text{NN-MARS}}(x)$ models the impact of each component x_i with a specific non-linear profile as illustrated in Figure 4. Overall, the NN-MARS is composed of $2d$ hidden neurons. Figure 1 shows that the hidden neurons operate in pairs. The training of NN-MARS is done with an ordinary gradient descent using the cross-entropy as loss function.

4 Experiments

We compare the performance and explainability of NN-MARS to decision trees (DTs), LR Natural Cubic Splines (LR NCS) [8, section 5.2] with fixed knots using uniform quantiles, MARS, and classical ReLU NNs. A 5-folds cross validation is realised for each tested method, such that the training sample is composed of 70% of the data and thus the validation set of the remaining 30%. All the experiments are implemented in Python with *scikit-learn* functionalities and *Pytorch* ones for the NNs. The computation times displayed in the article are reasonable since a computation server with a GPU is used.

Simulated data: We simulated $x \in \mathbb{R}^2$ data in order to visualize the estimated decision boundaries. We seek to predict the probability of developing a pathology as a function of cholesterol level (x_1) and weight (x_2). When the cholesterol level is low, the patient is more protected against the disease. On the other hand, when it is high, the patient is more at risk. Finally, being underweight or overweight increases the probability of developing the pathology. Since in real applications the boundary is noisy, we define the labels using a Bernouilli distribution with the estimated probability of developing the disease as parameter. The dataset is composed of 2000 patients, such that 1000 are sick (class 1).

Figure 2 presents the six methods compared. The red border represents the decision boundary of the tested method. The black lines represent the edges of the partitioning produced by the tested method. The deep NN (Fig. 2 - NN ($p = 30$)) is the best performing model because its estimated boundary is the closest to the simulated one. Nevertheless, the interpretation of its decision rule is too complex because the $\mathbb{R}^2$ space is partitioned into many polyhedra, some of which are not very useful for approximating the boundary. The DT, the LR NCS, MARS as well as NN-MARS segment the variables by hyperplanes making easier the interpretation of the decision rule (Fig. 2). Indeed, the partitioning depends on only one variable, contrary to the traditional NNs (Fig.2 - NN ($p = 4$) & NN($p = 30$)) that lead to oblique regions. For instance, with the proposed method, we can explain to the doctors that we have to take into account a 3 parts-segmentation of the cholesterol level : one for the patients with a level lower than -1.6, one for those with a value bigger than -0.3 and finally the intermediate ones. The MARS model with $M = 10$ (Fig.2 - MARS) defines 5 knots for the variable x_1 and 5 for x_2 in very close values. NN-MARS achieves the same performance as LR NCS with fewer spline functions, which justifies the need in automating the segmentation and not fixing the knots a priori.

The NN-MARS performs almost as well as the ReLU NNs while requiring fewer parameters to be estimated and ensuring a good explainability. Indeed, 9 parameters have to be optimized by the NN-MARS whereas the ReLU NNs with $p = 4$ and $p = 30$ require to estimate respectively 17 and 121 parameters. Although the NN-MARS estimates fewer parameters, the computation time required to converge to an optimal modeling is higher for the proposed method : $178(\pm1)$ seconds for one fold are required whereas the ReLU NNs with $p = 4$ and $p = 30$ take respectively $85(\pm26)$ and $120(\pm27)$ seconds. The standard deviations of computation time are higher for the traditional NNs because of the early stopping process that occur between 4000 and 8000 iterations at each fold, while the NN-MARS does not necessitate it and runs over 10000 epochs.

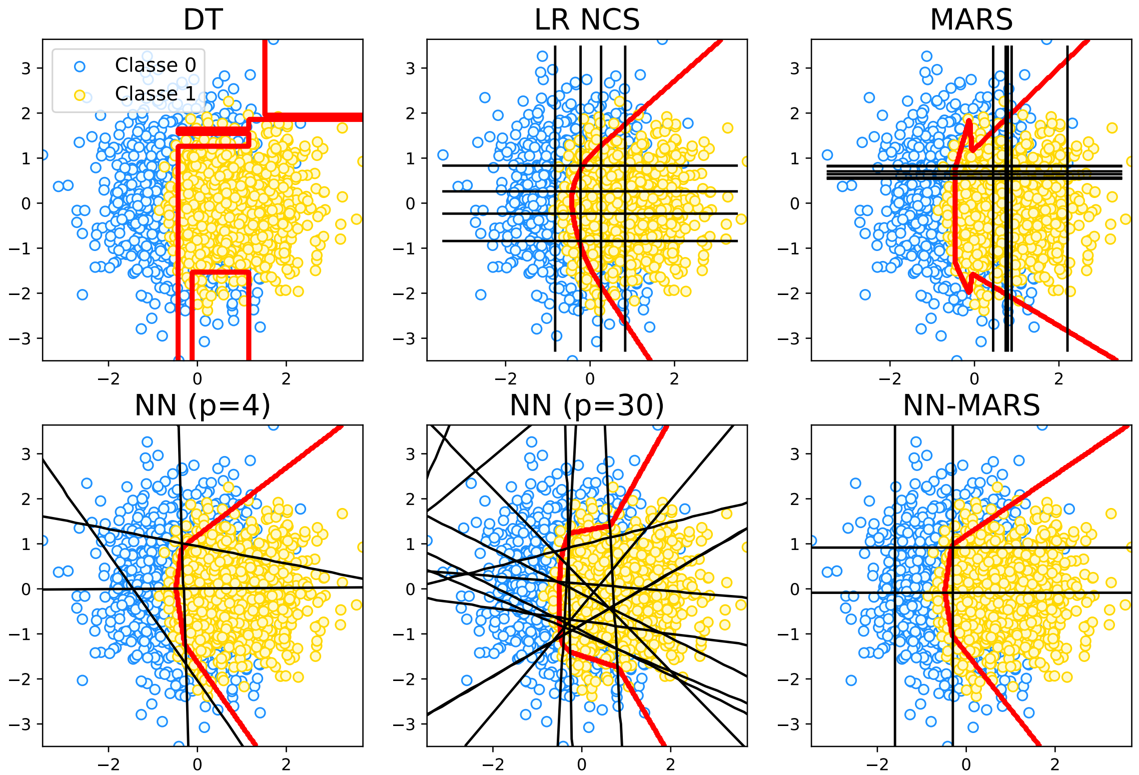


Fig. 2. Results on simulated data for DT, LR NCS, MARS model with $M = 10$, NN with $p = 4$ in (5), NN with $p = 30$ in (5), NN-MARS (4 nodes, since $d = 2$). Caption: estimated boundary in red, segmentation in black.

Finally, an experiment to test the scalability of the proposed method is presented on the figure 3. The AUC (Area Under the Curve) is a criterion appreciated by the doctors [9]. Where accuracy provides information about the predictive power of a model, the AUC gives an idea of the Sensibility and the Specificity of a predictive model, that is to say the capability of a tool for identifying a binary signal (sick or not in our case). The obtained average AUC over the 5-folds cross validation is displayed on both training and test samples for the NN-MARS, the ReLU NN with $p = 4$ and the ReLU NN with $p = 30$ according to different samplings. The proposed method (Fig. 3 curve in blue) obtains an AUC equivalent or even higher to the traditional ReLU NNs on the testing sample for all the number of patients tested. Indeed, traditional ReLU NNs tend to over-fit when only few patients are available in the training sample. The freezing of some parameters in the neural architecture of NN-MARS does not affect its predictive performance while improving the interpretability of its decision rule, even when a few number of patients is available in a cohort.

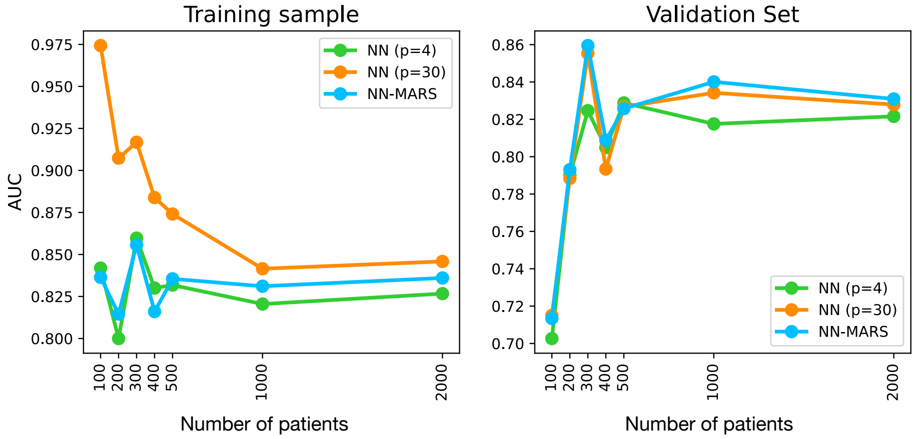


Fig. 3. AUC score (y-axis) on simulated data with different number of patients (x-axis) for the NN-MARS in blue, ReLU NN with $p = 4$ in green and ReLU NN with $p = 30$ in orange on training and validation samples.

Real data: We compared the performance of NN-MARS to other methods on a real data set. The main objective of the "Parkinson" database [10] is to detect people with Parkinson's disease from voice recordings of $N = 195$ patients ($24,6\%$ with Parkinson). We kept $d = 16$ biomedical measures of voice, such as maximum, average, and minimum voice frequencies. The NN-MARS is composed of 32 neurons. The training of the NNs is stopped when the error on the test data does not decrease anymore in order to avoid overfitting. The results are detailed in the table 1.

	Training set		Test set	
	Accuracy	AUC	Accuracy	AUC
LR	0.85 (0.02)	0.87 (0.02)	0.76 (0.01)	0.80 (0.06)
DT	0.91 (0.02)	0.94 (0.02)	0.88 (0.01)	0.77 (0.03)
LR NCS	0.90 (0.02)	0.94 (0.01)	0.82 (0.03)	0.87 (0.05)
MARS	0.90 (0.03)	0.91 (0.06)	0.82 (0.04)	0.89 (0.04)
NN ($p = 16$)	0.87 (0.04)	0.91 (0.07)	0.81 (0.06)	0.88 (0.07)
NN ($p = 70$)	0.86 (0.04)	0.91 (0.07)	0.83 (0.06)	0.88 (0.07)
NN-MARS	0.87 (0.01)	0.92 (0.03)	0.83 (0.05)	0.91 (0.05)

Tab. 1. Results of predictive performance on real data (mean and standard deviation in parentheses): the DT, the LR NCS, the 16-neurons NN, the 70-neurons NN, and the NN-MARS (6).

The LR is less efficient (76% accuracy on the test sample) than the other methods, which highlights the importance of introducing non-linearity in the modeling. Among all the non-linear methods tested, NN-MARS obtained the best AUC on the test sample (91%). The automation of the segmentation in MARS and NN models explains their higher AUCs than the LR NCS. The standard deviations show that the NN-MARS is more stable than the conventional NNs.

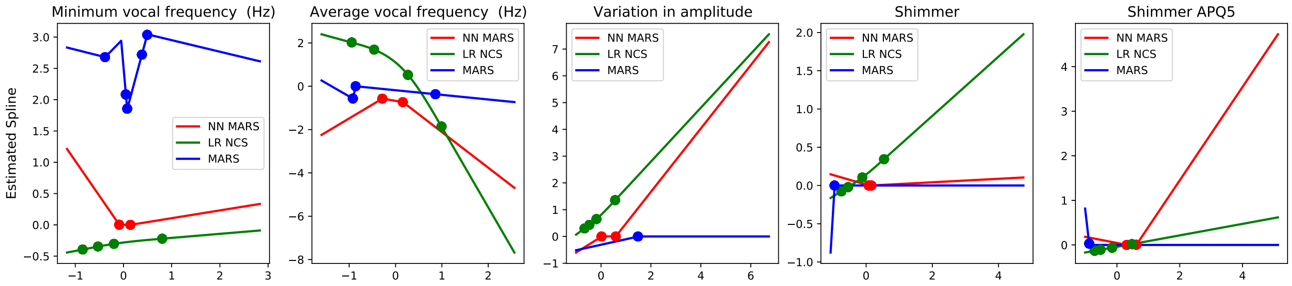


Fig. 4. Estimated splines on Parkinson dataset for the following features: Minimum voice frequency, Average voice frequency, Amplitude variation, Shimmer, Shimmer APQ5. Caption: NN-MARS in red, MARS in blue, LR NCS in green.

An advantage of the NN-MARS is the graphical interpretability of the results. Contrary to traditional ReLU NNs, the proposed method performs a partitioning with hyperplanes and thus the univariate estimated splines can be viewed. Figure 4 shows the estimated splines for 5 predictive variables. Each spline models the impact of the corresponding variable on the classification score. The estimated knots are represented by the points on the curves. The different methods find similar profiles for the variables but still quite dissimilar. For example, the NN-MARS (curves in red in Fig. 4) estimates that a low average vocal frequency increases the risks of developing Parkinson, since its estimated spline increases as the value of this feature decreases. From a certain level, the augmentation of the vocal frequency decreases the risk of being classified as sick. Moreover, this figure highlights the limits of the greedy learning of the MARS method. Only 5 variables are segmented, one of them 5 times (Fig. 4-Minimum vocal frequency). The MARS model is not able to increase its predictive performance by adding a new spline function, while it is possible to find a better performing classification rule, as demonstrated by the NNs in Table 1.

5 Conclusion

This paper develops an explainable neural network for nonlinear binary classification. Threshold effects are included in the modeling of the problem by approximating splines in the hidden layer of the network. The particular architecture of this network allows to control the segmentation of the variables and also to produce an easily understandable decision rule. Thus, this model is adapted to medical problems and to the expectations of specialists in this field. In future works, it would be interesting to include categorical variables and interactions between variables in order to improve predictive performance.

References

[1] Marie Guyomard, Dann J. Ouizeman, Renaud Schiappa, Cyprien Gilet, Jocelyn Gal, Emmanuel Chamorey, Stéphanie Patouraux, Thierry Piche, Albert Tran, Philippe Gual, Antonio Iannelli, Lionel Fillatre, and Rodolphe Anty. Diagnostic non-invasif de la nash fibrosante à l'aide de l'intelligence artificielle. *AFEF (Société Française d'Hépatologie)*, 2020.

[2] Douglas M Hawkins. On the choice of segments in piecewise approximation. *IMA Journal of Applied Mathematics*, 9(2):250–256, 1972.

[3] Asher Tishler and Israel Zang. A new maximum likelihood algorithm for piecewise regression. *Journal of the American Statistical Association*, 76(376):980–987, 1981.

[4] Konstantin Eckle et al. A comparison of deep networks with relu activation function and linear spline-type methods. *Neural Networks*, 110:232–242, 2019.

[5] Jerome H Friedman. Multivariate adaptive regression splines. *The annals of statistics*, 19(1):1–67, 1991.

[6] Ian J. Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep Learning*. MIT Press, Cambridge, MA, USA, 2016.

[7] Randall Balestriero et al. A spline theory of deep learning. In *International Conference on Machine Learning*, pages 374–383. PMLR, 2018.

[8] Trevor Hastie, Robert Tibshirani, Jerome H Friedman, and Jerome H Friedman. *The elements of statistical learning: data mining, inference, and prediction*, volume 2. Springer, 2009.

[9] Xiao-Hua Zhou, Donna K McClish, and Nancy A Obuchowski. *Statistical methods in diagnostic medicine*. John Wiley & Sons, 2009.

[10] Max Little et al. Suitability of dysphonia measurements for telemonitoring of parkinson's disease. *Nature Precedings*, 2008.

Session 3, Amphi A

Algorithms & data structures for sequences II

Genotyping multiscale variation with genome graphs

Brice Letcher[1], Martin Hunt[1] and Zamin Iqbal[1]
European Molecular Biology Laboratory, European Bioinformatics Institute (EMBL-EBI), Wellcome Genome Campus, Hinxton, Cambridgeshire, CB10 1SD, United Kingdom

Corresponding author: bletcher@ebi.ac.uk;zi@ebi.ac.uk

***Reference paper:* Brice Letcher, Martin Hunt, and Zamin Iqbal. Gramtools enables multiscale variation analysis with genome graphs. *Genome Biology*(2021).** https://doi.org/10.1186/s13059-021-02474-0

1 Introduction

Genome graphs are graph-based models capturing populations of genomes. Used in place of linear reference genomes, existing genome graph-based tools allow removing reference bias (mapped reads favouring the reference allele) [1] and mapping reads to highly diverged sequences, such as structural variants (SVs) [2]. When building graphs from large cohorts, however, genetic variation starts to occur at multiple length scales (e.g. SNPs and SVs) and on different references (e.g. SNPs on top of insertions), and no solutions currently exist to coherently model, genotype and output such multiscale variation. We developed a genome graph model and implementation to handle multiscale variation in the software `gramtools`.

2 Results

We model genome graphs as nested, directed, acyclic graphs (NDAGs). In NDAGs, variants can be contained inside (nested in) other variants and can occur on different references. We developed a model for genotyping variants in NDAGs in `gramtools`. As the standard genotyping output format VCF does not allow for alternate references, we also designed a JSON extension (jVCF) outputting all variants in the graph consistently.

We first validated genotyping with `gramtools` using simulated and real sequencing data at four highly variable genes in the malaria parasite *Plasmodium falciparum* (*Pf*). We then benchmarked `gramtools` against state-of-the art genome graph tools `vg` [1] and `GraphTyper2` [2] using a cohort of *M. tuberculosis* samples at regions with SNPs overlapping large deletions, finding superior genotyping performance. Finally, we applied `gramtools` to a case no existing tool can handle by genotyping a cohort of *Pf* samples at DBLMSP2, a gene known to contain two deeply diverged forms. Using `gramtools`, we could genotype and output variants segregating on top of each form, identifying previously unknown recombination between the two forms.

3 Conclusions

Genome graphs allow capturing a picture of dense, overlapping, and potentially multi-reference variation, but existing tools come short of coherently genotyping and outputting multiscale variation. We propose one solution in `gramtools` and demonstrate its applicability on two microbial pathogens. As cohorts continue to grow, so will this genotyping and representation problem, and we expect multiscale genotyping models and representation systems to grow in necessity and use.

References

[1] Erik Garrison, Jouni Sirén, Adam M Novak, Glenn Hickey, Jordan M Eizenga, Eric T Dawson, William Jones, Shilpa Garg, Charles Markello, Michael F Lin, Benedict Paten, and Richard Durbin. Variation graph toolkit improves read mapping by representing genetic variation in the reference. *Nature Biotechnology*, 36(9):875–879, August 2018.

[2] Hannes P. Eggertsson, Snaedis Kristmundsdottir, Doruk Beyter, Hakon Jonsson, Astros Skuladottir, Marteinn T. Hardarson, Daniel F. Gudbjartsson, Kari Stefansson, Bjarni V. Halldorsson, and Pall Melsted. GraphTyper2 enables population-scale genotyping of structural variation using pangenome graphs. *Nature Communications*, 10(1):5402, November 2019. Number: 1 Publisher: Nature Publishing Group.

Assessing conservation of alternative splicing with evolutionary splicing graphs

Diego Javier Zea[1], Elodie Laine[1] and Hugues Richard[2]

[1] Laboratory Computational and Quantitative Biology, IBPS, Sorbonne Université, 4 place Jussieu, 75005, Paris, France

[2] Bioinformatics Unit (MF1) Robert Koch Institute, Nordufer 20, 13353, Berlin, Germany

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

***Paper Reference:* Zea, DJ., Laskina, S., Baudin, A., Richard, H. and Laine E (2021) Assessing conservation of alternative splicing with evolutionary splicing graphs, Genome Research, 2021, 31(8):1462-1473. https://doi.org/10.1101/gr.274696.120**

Alternative splicing (AS) can significantly expand the proteome in eukaryotes by producing several transcript isoforms from the same gene. AS has been linked to morphological diversity, organ development, disease susceptibility, immune adaptation and interactome rewiring, among others. Although AS is well described at the genomic level, little is known about its contribution to protein evolution and the extent of the contribution of AS to proteome diversity has been a matter of debate. This question could be accurately addressed using evolutionary conservation. There is a clear need for computational methods that can couple the diversity of proteoforms resultings from AS with measures of sequence conservation across species.

Our work introduces a new method enabling for the first time granular estimates of alternative splicing conservation. It significantly improves our knowledge about the amount of functionally relevant variations. We first determine orthology relationships between exonic regions in the context of alternative splicing by extending Multiple Sequence Alignments to Splicing Graphs structure. We construct an Evolutionary Splicing Graph (ESG) where nodes define orthologous exon groups (denoted s-exons) and paths in the graph correspond to transcripts. The ESG summarizes the transcript variability observed across species, allowing the direct detection of conserved alternative splicing events. By analyzing AS conservation as far as teleosts, we show a clear link between the functional relevance, tissue-regulation and conservation of alternative splicing events on a set of 50 human genes. By constructing ESGs for the whole human proteome, we could annotate 46,000 evolutionary conserved AS events coming from 8,000 human protein-coding genes, dramatically changing previous estimates. We further identified a few thousands of genes where alternative splicing modulates the number and composition of pseudo-repeats shared across species.

The set of orthologous s-exons can be further analyzed to infer evolutionary scenarios explaining the observed transcripts variability. Each transcript history is described by a tree encapsulated in the tree of species, resulting in a phylogenetic forest. This representation allows us to understand the evolutionary relation between transcripts across species and date their age of creation. We have successfully applied it to study the evolution of the c-Jun N-terminal kinase (JNK) family [1].

Both works are made accessible as open source tools to the community. A web server (http://www.lcqb.upmc.fr/Ases), integrating both tools and providing interactive output is available [2]. It facilitates the study of alternative splicing evolution and its relation with the observed protein diversity by enabling interactive and linked representations of the evolutionary splicing graph and phylogenetic forest that were constructed.

References

1. Ait-Hamlat, A., Zea, DJ., Labeeuw, A., Polit, L., Richard, H. and Laine E. Transcripts' Evolutionary History and Structural Dynamics Give Mechanistic Insights into the Functional Diversity of the JNK Family". JMB 432.7 (2020), pp. 2121–2140
2. Zea., DJ., Richard, H. and Laine E."ASES: Visualizing evolutionary conservation of alternative splicing in proteins". Bioinformatics (2022).

Session 3, Amphi B

Structural bioinformatics & proteomics I

Conformational variability in proteins bound to single-stranded DNA: A new benchmark for new docking perspectives

Dominique Mias-Lucquin[1] and Isaure Chauvot de Beauchene[1,2]

[1] LORIA, Universite de Lorraine, 54500 Vandœuvre-lès-Nancy, France.
[2] CNRS, 54500 Vandœuvre-lès-Nancy, France.

Corresponding Author: dominique.mias-lucquin@loria.fr

Paper Reference: **Mias-Lucquin *et al.* (2018) Conformational variability in proteins bound to single-stranded DNA: A new benchmark for new docking perspectives, Proteins, 2022 Mar;90(3):625-631. https://doi.org/10.1002/prot.26258**

While originally described by Watson and Crick as a double helix, composed of two strands bonded together by hydrogen bonds, DNA is often found in a transient single-stranded state (ssDNA) during its processing, such as genome replication,or horizontal gene transfer,and bound to proteins. These complexes (ribosomes, ICE-relaxase, replication fork complex and so forth) are potential therapeutic targets in diseases.

The structural analysis of these complexes can help to understand how they achieve their function. For example, it can reveal the conformational changes undergone by the protein during nucleic acids (NAs) binding, by comparing protein structures with and without bound NA[1].

While very informative, high-resolution experimental structures of ssNA–protein complexes are expensive and may be difficult, or even impossible, to obtain, due to the inherent poor ordering of NA, especially ssNA. Several software systems have tried to implement accurate ssRNA–protein docking (ATTRACT[2], RNP-denovo[3], RNA-lim[4]).

While all these methods advertise a prediction precision from 2 to 10 Å of RMSD between predicted and experimental ssRNA location, none of them was tested yet on ssDNA–protein docking. To our knowledge, no benchmark is available for ssDNA–protein docking. Moreover, while it is possible to query ssDNA–protein complexes with the Nucleic Acid Database (NDB), it seems to find none after 2013, thus limiting the scope of a NDB-derived benchmark.

In turn, docking algorithms need experimental ground truth to validate and compare methods. Such docking benchmarks exist for protein-protein, membrane protein-protein, protein-RNA and dsDNA-protein complexes. Also, while some works studied ssDNA–protein interactions from few structures in the Protein Data Bank (PDB), none seems to be as exhaustive as possible, with a primary goal to improve ssDNA–protein docking.

Here, we present an ssDNA–protein docking benchmark based on structures extracted from the PDB that contains 91 sequence-identity groups of bound–unbound protein chains, created to evaluate ssDNA–protein docking. Due to the high flexibility of unstructured ssNA, it is not relevant to use their unbound forms in the context of macromolecular docking. This is also the reason why the docking programs presented earlier do not require a known unbound ssRNA structure. In consequence, the main aim of this dataset is to provide bound and unbound structures of the proteins but only bound structures of ssDNA, from ssDNA–protein complexes. When possible, we provide several structures for both bound and unbound states, allowing to differentiate binding-specific from binding-independent conformational changes.

Docking experiments were performed to show a use case for this benchmark. It underlines the relevance of using several bound structures as ground truth and to tolerate a minimum conformational deviation from ground truth when evaluating docking results.

References

[1] Büttner K, Nehring S, Hopfner K-P. Structural basis for DNA duplex separation by a superfamily-2 helicase. Nat Struct Mol Biol. 2007; 14(7): 647- 652. https://doi.org/10.1038/nsmb1246

[2] Chauvot de Beauchene I, de Vries SJ, Zacharias M. Binding site identification and flexible docking of single stranded RNA to proteins using a fragment-based approach. PLoS Comput Biol. 2016; 12(1) :e1004697. https://doi.org/10.1371/journal.pcbi.1004697

[3] Kappel K, Das R. Sampling native-like structures of RNA-protein complexes through Rosetta folding and docking. Structure. 2019; 27(1): 140- 151.e5. https://doi.org/10.1016/j.str.2018.10.001

[4] Hall D, Li S, Yamashita K, Azuma R, Carver JA, Standley DM. RNA-LIM: a novel procedure for analyzing protein/single-stranded RNA propensity data with concomitant estimation of interface structure. Anal Biochem. 2015; 472: 52- 61. https://doi.org/10.1016/j.ab.2014.11.004

Analysis of SARS-CoV-2 and human protein interactions: a CAPRI-COVID Round

Théo Mauri[1], Guillaume Brysbaert[1], Paul A. Bates[2], Shoshana Wodak[3] and Marc F. Lensink[1]
[1] Université de Lille, CNRS, UMR 8576-UGSF-Unité de Glycobiologie Structurale et Fonctionnelle, F-59000 Lille
[2] Biomolecular Modelling Laboratory, The Francis Crick Institute, London, UK.
[3] Center for Structural Biology, VIB-VUB, Brussels, Belgium

Corresponding author: theo.mauri@univ-lille.fr

Abstract *In the beginning of the CoViD-19 pandemic, high-throughput AP-MS experiments established putative interaction maps between SARS-CoV-2 and human proteins. shortly thereafter, the CAPRI committee selected 4 complexes from these for prediction of their three-dimensional coordinates in Round 51. A fifth target - having no resolved structure - was added coming from the previous joint CASP14-CAPRI prediction. The CAPRI community submitted models of interaction but, as no experimental structures are available so far, the quality of the predicted structures of each target is therefore difficult to assess. Instead, a novel alternative protocol was used to evaluate the prediction results. This protocol was developed based on an initial consultation with participants of Round 51, and was further defined by a working group. This presentation will highlight the different methods used to find that consensus.*
Firstly, the propensity of residues in each protein (receptor and ligand) to contribute to the interaction was investigated. Secondly, the frequency with which protein residues engage in inter-molecular contacts in the predicted models, is compared to residue conservation levels, which tend to indicate involvement in biological function. These results have been presented at the CAPRI community exchange portal (https://www.ebi.ac.uk/pdbe/complex-pred/capri/round/51/).
The consensus methods must be validated on verified data. To this, different targets from the CAPRI Score_set 2014 have been selected. As the difficulty of the SARS-CoV-2 target is unknown, we selected from this score_set different difficulties according to the assessments. Thus, in a third part, the methods tested on the evaluated target will be apply to the CAPRI COVID Round to find potential consensus answers.

In this study we developed a new method to assess models quality without template.

Keywords CoViD-19, CAPRI, Protein complexes, Clustering, Protein-protein interaction

1 Introduction

Early 2020 is the beginning of a worldwide pandemic due to the virulence of the SARS-CoV-2. To help combat this sanitary crisis, the scientific community provides a lot of effort to understand the different mechanisms involved in the infection by this virus. A study of Gordon *et al.* (2020) provides a set of 332 high-confidence SARS-CoV-2-human protein-protein interactions (PPIs)[1], using affinity-purification–mass spectrometry (AP-MS) on HEK-293T/17 cell line infected with the SARS-CoV-2 virus.

From these results, the CAPRI (Critical Assessment of PRotein Interactions) committee selected interactions for community prediction, based on confidence level and availability of modeling templates. These four complexes called targets (T182, T183, T184 and T185) have been proposed to the CAPRI community in the aim to model these interactions and have a better understanding of the virus mechanisms [2]. In addition, another target (T181) was added from the joint CASP14-CAPRI prediction as it is a virus-host complex unsolved experimentally [3].

For each target, about thirty predictor groups participated producing between 1250 and 2000 models. These models have been made available to the community as part of the CAPRI COVID-19 Open Science initiative. From those, nineteen scorers groups have selected less than 200 models.

Here are described the different models selected by the scorers and the methods used to define a potential consensus for each target. To that, we investigated the proportion of residues involved in the interaction but also specific contacts. Results have been analyzed through clustering and meta-clustering and end with a new and objective method to define best models according to all the different scorers groups.

2 Material and Methods

2.1 Target presentations

In this study, we focused on 4 targets (T181 to T184) of CoViD-19 CAPRI prediction Round. The number of models submitted by the predictor and scorer groups are listed in the Tab.1. Further information about the targets are given in the Tab.2 and on this website.

Target ID	Predictors Groups	Predictors Models	Scorers Groups	Scorers Models
T181	26	1257	19	185
T182	30	1972	19	181
T183	27	1523	19	164
T184	30	1811	19	190

Tab. 1. Summary of target participation; number of models submitted and evaluated.

CAPRI ID	CASP ID	Components	Uniprot IDs	PDB Templates
T181	H1103	Orf3a / HMOX1	P0DTC3 / P09601	6XDC / 1N3U
T182	NA	Nsp15 / NUTF2	P0DTD1 / P61970	6WLC / 1GY5
T183	NA	Nsp8 / EXOSC8	P0DTD1 / Q96B26	2NN6 / 3UB0:D, 2AHM:G, 6XIP
T184	NA	Nsp7 / RhoA	P0DTD1 / P61586	6XIP:C, 3UB0:C, 6M5I:A / 5C2K:A, 4LHW:E, 2J1L:A190

Tab. 2. Summary of target participation.

2.2 Validation dataset

We selected 4 targets from previous CAPRI Rounds where the different models have been assessed to different quality according to the CAPRI assessment criteria: f_{nat} and $f_{non-nat}$ which are the fraction of receptor-ligand residue contacts found in the model and the experimental structure and the fraction of contacts which have been predicted in the model and which are not present in the target structure[4]. In addition, to assess the quality of the predicted interface, two other quantities based on Root Mean Square Distance (RMSD) are used: I-rms which is the rmsd at the interface backbone atoms and S-rms which is the same for the interface side-chain atoms. The different qualities are the following: incorrect, acceptable, medium and high.

The selected targets are listed below with their subjective difficulty according to the number of acceptable or above provide by their scorer sets:

— T039: considered here as a hard complex to predict
— T041: considered here as an easy complex to predict
— T050: considered as a medium difficulty target to predict with a A1B2 stoichiometry
— T053: considered as a medium difficulty target to predict

The number of models and the quality of the different targets are summarized in the following table (Tab.3):

Target ID	Number of models	incorrect models	acceptable models	medium models	high models
T039	120	120	0	0	0
T041	120	46	54	18	2
T050	140	105	21	14	0
T053	130	92	25	13	0

Tab. 3. Summary of validation targets regarding the number of models and the assessment quality

Also, to get a larger amount of data to test the method developed in this study, we used the updated scoreset of CAPRI[4]. This new version is available at scoreset.org and the related article is being written. In total, 120 targets with an amount of 121209 models have been analyzed. Most of these targets can be divided into 4 different kingdoms: Archaea, Bacteria, Viruses and Eukaryotes.

The other targets are interaction between two different kingdoms (Eukaryotes and Bacteria) or are unknown. The distribution of the number of targets and models and the percentage are listed in the Tab.4.

Kingdom	Number of targets	Number of models	Percentage of the total set
Archaea	18	7890	6.51%
Bacteria	47	45908	37.88%
Eukaryotes	41	50645	41.78%
Viruses	10	7848	6.47%
Other	4	8918	7.36%

Tab. 4. Proportion of new scoreset targets in the four different kingdoms

2.3 Interface residue composition

To compare the residue composition at the interface for the different models, we transformed all models into RINs (Residue Interaction Networks). These networks establish an edge (interaction) between pair residues if their distance falls between 1.5 and 5Å[5,6]. From this information, the residues at the interface are retrieved and then the number of times a residue has been predicted to be at the interface in the different models calculated. This method has also been used to retrieve the specific contacts (meaning two specific residues interacting) between residues from different components.

2.4 Residue conservation

To add structural visualization for better understanding, we decided to color the contact residues regarding their conservation. To compute this conservation, we used the Rate4Site tool. This software is based on the maximum likelihood principle and maps the rate of evolution among homologous proteins onto the surface of the molecule of one of the homologous proteins with a known three-dimensional structure [7]. A rate is calculated for each residue of the protein sequence and can be between 0 and 1 where 0 corresponds to a residue not conserved and 1 well conserved among homologous proteins.

2.5 Visualization

To easily compare residue conservation and the one predicted to be in contact, molecular visualization is a powerful tool. For each target, a PyMOL session has been set with the two same molecules in the same orientation [8]. For one molecule, the bfactor of the pdb files have been replaced with the conservation and the other molecule with the number of times the residue have been predicted to be in contact normalized by the number of models.

Also, to visualize where the location of the interactions are, for every model, the center of mass of the different elements have been calculated. These coordinates have been set to a sphere visible.

2.6 Clustering

To find a consensus between all the models, we performed a clustering on all the models based on different features: interface residues (residues with are up to 10Åfrom the other chain of model), contact residues (up to 5Å) and specific contacts. As clustering is a method to see the distribution of elements according to their distances, we defined the distance (D) between two models as followed:

— Interface and contact residues:
$D = 1 - \frac{R1 \bigcap R2}{R1 \bigcup R2}$ with $R1$ interface or contact Residue model1 and $R2$ from model2

— Specific contacts:
$D = 1 - \frac{SC1 \bigcap SC2}{SC1 \bigcup SC2}$ with $SC1$ specific contacts from model1 and $SC2$ from model2

If two different models have a distance of 0, it is considered as redundant because it can be a model selected by two different scorers. In that case the second one is removed from the list of models.

To this we used the R software with agnes (Agglomerative Nesting) R package to be able to cluster the solutions.

Different clustering methods have been compared with the linkage criteria: Average, Single, Complete and Ward. In principle, Hierarchical clustering produces an amount of clusters starting with the number of elements and decreasing iteratively to one big cluster [9]. To set a representative number of clusters to represent every different solution, we decided to choose a cut-off value. This value corresponds to the similarity inside a cluster, so a cut-off of 0.25 means that inside a cluster every models have at least 75% of similarity or a distance $\leqslant$ 0.25 pair-wised. In this study we selected a cut-off of 0.25 after trying different ones like 0.4 and 0.6

2.7 Meta-clustering

To have a better point of view of the results, we decided to perform some meta-clustering with Cytoscape using Markov clustering (MCL) based on the Jaccard index (1-Distance) between the most representative structure of each cluster (medoid)[10,11]. This allows us to create bigger clusters where the clusters are closer. For the validation set, this will allow us to see if the biggest meta-cluster contains the best solution.

2.8 Adjacency overlap

To see how well a model matches the consensus of all models, we calculated the adjacency overlap. For each target an adjacency matrix was constructed from the interaction network. If a specific interaction is detected between two residues in a model, this very contact is incremented by 1. Once the matrix is set with all the models, the values are normalized with the number of models in total and then every model is compared to this matrix to see how well this model matches the matrix.

3 Results

3.1 Score_set targets

Clustering and meta-clustering

To select the hierarchical clustering, the agglomerative coefficient of the four methods have been tested and are retrieved in the Tab.5. Regarding these scores, the Ward method is the one that has to be used for the following results.

Target ID	Average	Single	Complete	Ward
T181	0.8006	0.7738	0.8377	0.9365
T182	0.9398	0.9115	0.9571	0.9964
T183	0.9191	0.8801	0.9555	0.9947
T184	0.8916	0.8023	0.9301	0.9945

Tab. 5. Agglomerative coefficient of the four hierarchical clustering methods (”Average”, ”Single”, ”Complete”, ”Ward”) for the fourth targets T181, T182, T183 and T184

After analyzing the clustering based on specific contacts,the composition of contact and interface residues, we choose to focus only on specific contacts to have a more fine grain vision.

With a threshold of 0.25, meaning that the models inside a cluster have at least 75% similarity, the number of clusters depends on each target. Regarding T039, considered a hard target as there is no correct model in the models proposed by the scorers, the number of clusters regarding contact residues is 108 and the biggest one contains 4 models coming from 4 different scorers. The easy target T041 has all its models collected in 105 clusters, a maximum size of three found twice. T050's models, with a A1B2 stoichiometry, have been grouped into 113 clusters with 6 six clusters with the biggest size of 3 models. The models of T053, an intermediate target regarding the difficulty, are gathered into 107 clusters with a maximum size of 3 for three clusters.

As the number of clusters is important regardless of the difficulty of the different targets, we decided to perform a meta-clustering to see if the biggest meta-cluster contains the good solution.

For an easy target as T041 we can see on Fig.1(**A**) that all the good solutions are in the same meta cluster and are well connected but in a difficult target (T039) Fig.1 (**B**), there is a lot of meta-clusters with low amount of link. For the two other targets, good results are gathered in the first meta-cluster. For T053, incorrect models are also found in the biggest meta-cluster. It could be interesting to find a method to remove them.

Adjacency overlap

Looking how well a model matches inside a consensus could be a way to discriminate good from bad models. To that we performed a new method on our testing dataset T039, T041, T050 and T053 called Adjacency overlap.

According to the different plots on Fig.2, we can see that models with a good prediction have a higher score than incorrect models even if some ”high” quality models don’t have the highest score. Regarding the adjacency overlap score, it seems to have a threshold value of 0.04 above which models are often correct. For some Targets such as T053 and T050 acceptable models are found under this threshold. For T039 all the models are below this threshold but the one which is close has a f_{nat} near the ”correct” assessment criteria.

3.2 CAPRI COVID Targets

Comparison of the conservation and the frequency of residues

To know if residues well conserved in the human and the SARS-CoV-2 proteins are often predicted to be in contact, residues around $5\AA$ of the interaction interface have been counted and retrieved. Regarding the surface of the two molecules (human and virus), residues of the human proteins which are predicted to be in contact with the virus are often well conserved contrarily to the viral proteins. Such results can be observed on Fig.3. With the spheres around the molecules we can see the spread of the ligand positions for every model prediction. This distribution shows a low tendency about a certain location, but most ligand predicted positions are all around the receptor.

These results can be found on the capri-docking webpage (https://www.capri-docking.org//capri-covid/).

Clustering and meta clustering

For every target, the number of clusters is close to the number of models meaning that models are not similar while looking at specific contacts between the different components. The number of clusters for T181 is 120, 119 for T182, 102 for T183 and 115 for T184. The biggest clusters have a size up to 4 models. These results are not sufficient to easily identify a good solution but we can use the same approach as with the validation dataset. Thus, we also performed a meta-clustering to see if there is an interesting meta-cluster. The results are similar to the T039 ones with a high number of meta-clusters and light links between them meaning there is no consensus yet.

Adjacency overlap

Looking at the different curves from Fig.4, we can see that all the models fall below the defined threshold except for two models of the T184 (Fig.4 (**C**)). Some models tend to have a better score than others and it could be interesting to focus on them. At first sight, the ranking of the different scorers seems to be non relevant regarding this method because of the heterogeneous results. But, as all scorers are mixed up, it could be interesting to see results of scorers independently to see if the two methods can be combined. Regarding the adjacency overlap results we can assume that some targets are easier than others like T182 and T184.

We tested the method we developed on the new CAPRI scoreset to see if this method is a good one to discriminate right from wrong models. A Receiver Operating Characteristic (ROC) curve and a precision-recall curve have been calculated regarding the all set but also to the specific kingdom of our study in Fig. The figure 1.A shows the overall capacity of the new scoring method to discriminate right models from wrong models with an Area Under Curve (AUC) of 0.905. But these results are regardless of the kingdom. In this study, we are looking for the interaction between human and virus proteins. SO we looked at the ROC curves for those specific kingdoms. According to the eukaryotes kingdom (Fig5.2.A), the AUC of 0.880 is lower than the full set but the results are still good which is better than the virus set (Fig5.3.A) which have an AUC of 0.758. To see the precision of this method, the precision-recall curves have been plotted. These curves allow us to see the precision of this method while increasing the sensitivity. We can see that we can have a good precision for the whole set and the Eukaryote one but with a low sensitivity meaning that some good models will be missed. For

Viruses the precision is very low. We can use these curves to define a threshold to maximize the TPR and the precision. These results are summarized in the Tab.6 and 7.

Set	Threshold	TPR	FPR
full set	0.0183	0.9849	0.3600
Eukaryotes	0.0258	0.9664	0.3200
Viruses	0.0285	0.9646	0.35

Tab. 6. Threshold of the scoring method and the True Positive Rate (TPR) and False Positive Rate (FPR) associated for the full scoreset, the eukaryote and viruses ones.

Set	Threshold	Precision	Sensitivity
full set	0.0848	0.8002	0.2750
Eukaryotes	0.1007	0.9331	0.0373
Viruses	0.0319	0.3015	0.9407

Tab. 7. Threshold of the scoring method and the precision and the sensitivity associated for the full set, the eukaryote set and viruses set.

4 Discussion - Conclusion

Looking at the correlation between conservation and presence at the interface shows us two different things: human residues predicted to be at the interface are well conserved which is not the case for the viral protein. From these results we can hypothesize that the virus residues at the interface change to adapt and infect many organisms.

Testing methods on a heterogeneous validation dataset allow us to obtain some patterns. Clustering regarding the specific contacts only regroups very close models, the meta-clustering brings more information but it is not enough to find a consensus. The adjacency overlap seems to allow us to remove incorrect models from the meta clustering and select good results. This method applied to the full scoreset of CAPRI, gave better results than expected with the four targets with a lower threshold. But as the complexes are between two different species it could be more effective to focus on the most stringent thresholds. If we want to optimize the precision, a threshold of 0.0319 is better. With this, the probability of having good models is only 30.15%. But only two targets have models above this threshold: T182 and T184.

The different methods applied to the complexes between human and viral proteins did not lead to a concrete solution. This can be explained by the complexity to predict a model between two different organisms, especially when viruses are involved according to the scoreset. Also, the study where each target has been predicted to interact has been made through affinity-purification mass spectrometry but the specific interaction between the different components is not one hundred percent sure and the hypothesis of other proteins involved in the interaction is still present.

Acknowledgements

The authors express gratitude to the ensemble of CAPRI predictor groups who participate in this special Round.

References

[1] David E. Gordon, Gwendolyn M. Jang, Mehdi Bouhaddou, Jiewei Xu, Kirsten Obernier, Kris M. White, Matthew J. O'Meara, Veronica V. Rezelj, Jeffrey Z. Guo, Danielle L. Swaney, Tia A. Tummino, Ruth Hüttenhain, Robyn M. Kaake, Alicia L. Richards, Beril Tutuncuoglu, Helene Foussard, Jyoti Batra, Kelsey Haas, Maya Modak, Minkyu Kim, Paige Haas, Benjamin J. Polacco, Hannes Braberg, Jacqueline M. Fabius, Manon Eckhardt, Margaret Soucheray, Melanie J. Bennett, Merve Cakir, Michael J. McGregor, Qiongyu Li, Bjoern Meyer, Ferdinand Roesch, Thomas Vallet, Alice Mac Kain, Lisa Miorin, Elena Moreno, Zun Zar Chi Naing, Yuan Zhou, Shiming Peng, Ying Shi, Ziyang Zhang, Wenqi Shen, Ilsa T. Kirby, James E. Melnyk, John S. Chorba, Kevin Lou, Shizhong A. Dai, Inigo Barrio-Hernandez, Danish Memon, Claudia Hernandez-Armenta, Jiankun Lyu, Christopher J. P. Mathy, Tina Perica, Kala Bharath Pilla, Sai J. Ganesan, Daniel J. Saltzberg, Ramachandran Rakesh, Xi Liu, Sara B. Rosenthal, Lorenzo Calviello, Srivats Venkataramanan, Jose Liboy-Lugo, Yizhu Lin, Xi-Ping Huang, YongFeng Liu, Stephanie A. Wankowicz, Markus Bohn, Maliheh Safari, Fatima S. Ugur, Cassandra Koh, Nastaran Sadat Savar, Quang Dinh Tran,

Djoshkun Shengjuler, Sabrina J. Fletcher, Michael C. O'Neal, Yiming Cai, Jason C. J. Chang, David J. Broadhurst, Saker Klippsten, Phillip P. Sharp, Nicole A. Wenzell, Duygu Kuzuoglu-Ozturk, Hao-Yuan Wang, Raphael Trenker, Janet M. Young, Devin A. Cavero, Joseph Hiatt, Theodore L. Roth, Ujjwal Rathore, Advait Subramanian, Julia Noack, Mathieu Hubert, Robert M. Stroud, Alan D. Frankel, Oren S. Rosenberg, Kliment A. Verba, David A. Agard, Melanie Ott, Michael Emerman, Natalia Jura, Mark von Zastrow, Eric Verdin, Alan Ashworth, Olivier Schwartz, Christophe d'Enfert, Shaeri Mukherjee, Matt Jacobson, Harmit S. Malik, Danica G. Fujimori, Trey Ideker, Charles S. Craik, Stephen N. Floor, James S. Fraser, John D. Gross, Andrej Sali, Bryan L. Roth, Davide Ruggero, Jack Taunton, Tanja Kortemme, Pedro Beltrao, Marco Vignuzzi, Adolfo García-Sastre, Kevan M. Shokat, Brian K. Shoichet, and Nevan J. Krogan. A SARS-CoV-2 protein interaction map reveals targets for drug repurposing. *Nature*, 583(7816):459–468, July 2020.

[2] Marc F. Lensink and Shoshana J. Wodak. Docking, scoring, and affinity prediction in CAPRI. *Proteins*, 81(12):2082–2095, December 2013.

[3] Marc F. Lensink, Guillaume Brysbaert, Théo Mauri, Nurul Nadzirin, Sameer Velankar, Raphael A. G. Chaleil, Tereza Clarence, Paul A. Bates, Ren Kong, Bin Liu, Guangbo Yang, Ming Liu, Hang Shi, Xufeng Lu, Shan Chang, Raj S. Roy, Farhan Quadir, Jian Liu, Jianlin Cheng, Anna Antoniak, Cezary Czaplewski, Artur Giełdoń, Mateusz Kogut, Agnieszka G. Lipska, Adam Liwo, Emilia A. Lubecka, Martyna Maszota-Zieleniak, Adam K. Sieradzan, Rafał Ślusarz, Patryk A. Wesołowski, Karolina Zieba, Carlos A. Del Carpio Muñoz, Eiichiro Ichiishi, Ameya Harmalkar, Jeffrey J. Gray, Alexandre M. J. J. Bonvin, Francesco Ambrosetti, Rodrigo Vargas Honorato, Zuzana Jandova, Brian Jiménez-García, Panagiotis I. Koukos, Siri Van Keulen, Charlotte W. Van Noort, Manon Réau, Jorge Roel-Touris, Sergei Kotelnikov, Dzmitry Padhorny, Kathryn A. Porter, Andrey Alekseenko, Mikhail Ignatov, Israel Desta, Ryota Ashizawa, Zhuyezi Sun, Usman Ghani, Nasser Hashemi, Sandor Vajda, Dima Kozakov, Mireia Rosell, Luis A. Rodríguez-Lumbreras, Juan Fernandez-Recio, Agnieszka Karczynska, Sergei Grudinin, Yumeng Yan, Hao Li, Peicong Lin, Sheng-You Huang, Charles Christoffer, Genki Terashi, Jacob Verburgt, Daipayan Sarkar, Tunde Aderinwale, Xiao Wang, Daisuke Kihara, Tsukasa Nakamura, Yuya Hanazono, Ragul Gowthaman, Johnathan D. Guest, Rui Yin, Ghazaleh Taherzadeh, Brian G. Pierce, Didier Barradas-Bautista, Zhen Cao, Luigi Cavallo, Romina Oliva, Yuanfei Sun, Shaowen Zhu, Yang Shen, Taeyong Park, Hyeonuk Woo, Jinsol Yang, Sohee Kwon, Jonghun Won, Chaok Seok, Yasuomi Kiyota, Shinpei Kobayashi, Yoshiki Harada, Mayuko Takeda-Shitaka, Petras J. Kundrotas, Amar Singh, Ilya A. Vakser, Justas Dapkūnas, Kliment Olechnovič, Česlovas Venclovas, Rui Duan, Liming Qiu, Xianjin Xu, Shuang Zhang, Xiaoqin Zou, and Shoshana J. Wodak. Prediction of protein assemblies, the next frontier: The CASP14-CAPRI experiment. *Proteins*, 89(12):1800–1823, December 2021.

[4] Marc F. Lensink and Shoshana J. Wodak. Score_set: a CAPRI benchmark for scoring protein complexes. *Proteins*, 82(11):3163–3169, November 2014.

[5] Guillaume Brysbaert, Théo Mauri, and Marc F. Lensink. Comparing protein structures with RINspector automation in Cytoscape. *F1000Res*, 7:563, 2018.

[6] Gil Amitai, Arye Shemesh, Einat Sitbon, Maxim Shklar, Dvir Netanely, Ilya Venger, and Shmuel Pietrokovski. Network analysis of protein structures identifies functional residues. *J Mol Biol*, 344(4):1135–1146, December 2004.

[7] Tal Pupko, Rachel E. Bell, Itay Mayrose, Fabian Glaser, and Nir Ben-Tal. Rate4Site: an algorithmic tool for the identification of functional regions in proteins by surface mapping of evolutionary determinants within their homologues. *Bioinformatics*, 18 Suppl 1:S71–77, 2002.

[8] DeLano, W.L. The PyMOL Molecular Graphics System, 2002.

[9] G. J. McLachlan, R. W. Bean, and S. K. Ng. Clustering. In Jonathan M. Keith, editor, *Bioinformatics*, volume 1526, pages 345–362. Springer New York, New York, NY, 2017. Series Title: Methods in Molecular Biology.

[10] Paul Shannon, Andrew Markiel, Owen Ozier, Nitin S. Baliga, Jonathan T. Wang, Daniel Ramage, Nada Amin, Benno Schwikowski, and Trey Ideker. Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res*, 13(11):2498–2504, November 2003.

[11] James Vlasblom and Shoshana J. Wodak. Markov clustering versus affinity propagation for the partitioning of protein interaction graphs. *BMC Bioinformatics*, 10:99, March 2009.

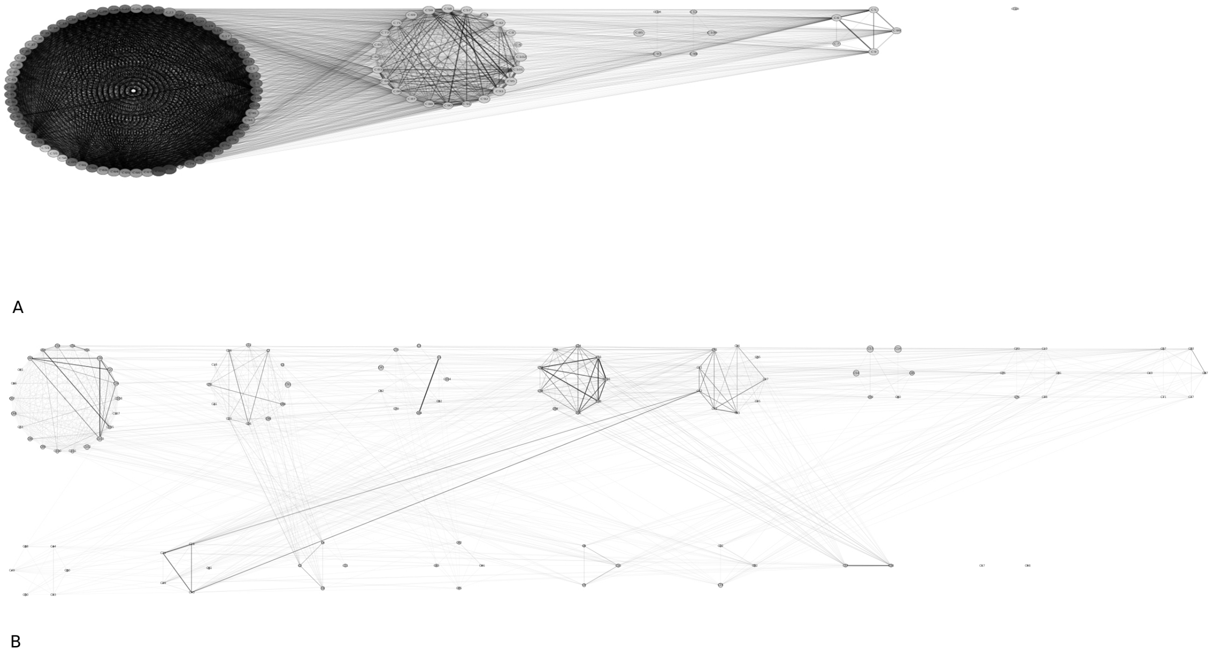


Fig. 1. Meta-clustering of T041 (**A**) and T039 (**B**). The colour reflect the quality of the models inside the cluster from light grey meaning that models are incorrect to dark grey where models quality are high. The size of the spheres corresponds to the number of models inside the cluster. Links correspond to Jaccard index: strong links mean strong similarity

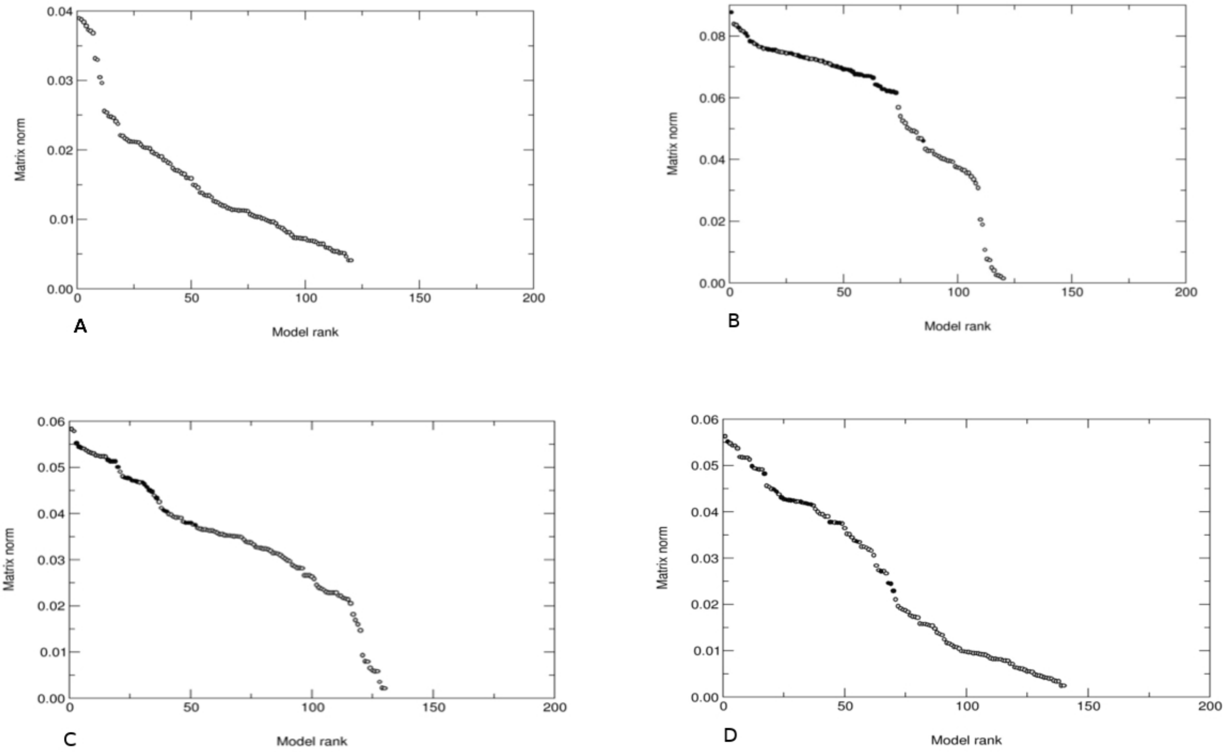


Fig. 2. Adjacency overlap results of T039 (**A**), T041 (**B**), T050 (**C**) and T053 (**D**). The colour reflects the quality of the models from light grey meaning that models are incorrect to dark grey where models quality are high.

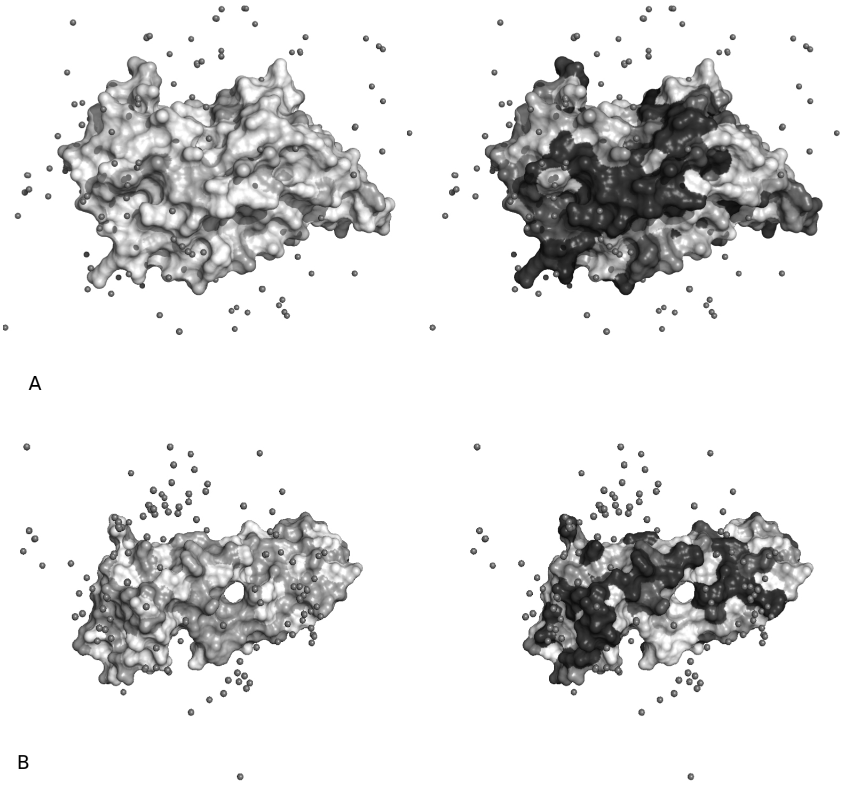


Fig. 3. Surface representation of the component of T181. (**A**) is the representation of the human protein and (**B**) is the virus visualisation. Left corresponds to the residue conservation from light grey to dark grey and Right to the contact hits where dark grey is often hit. Spheres around are the centers of mass of every partner predicted by the different models.

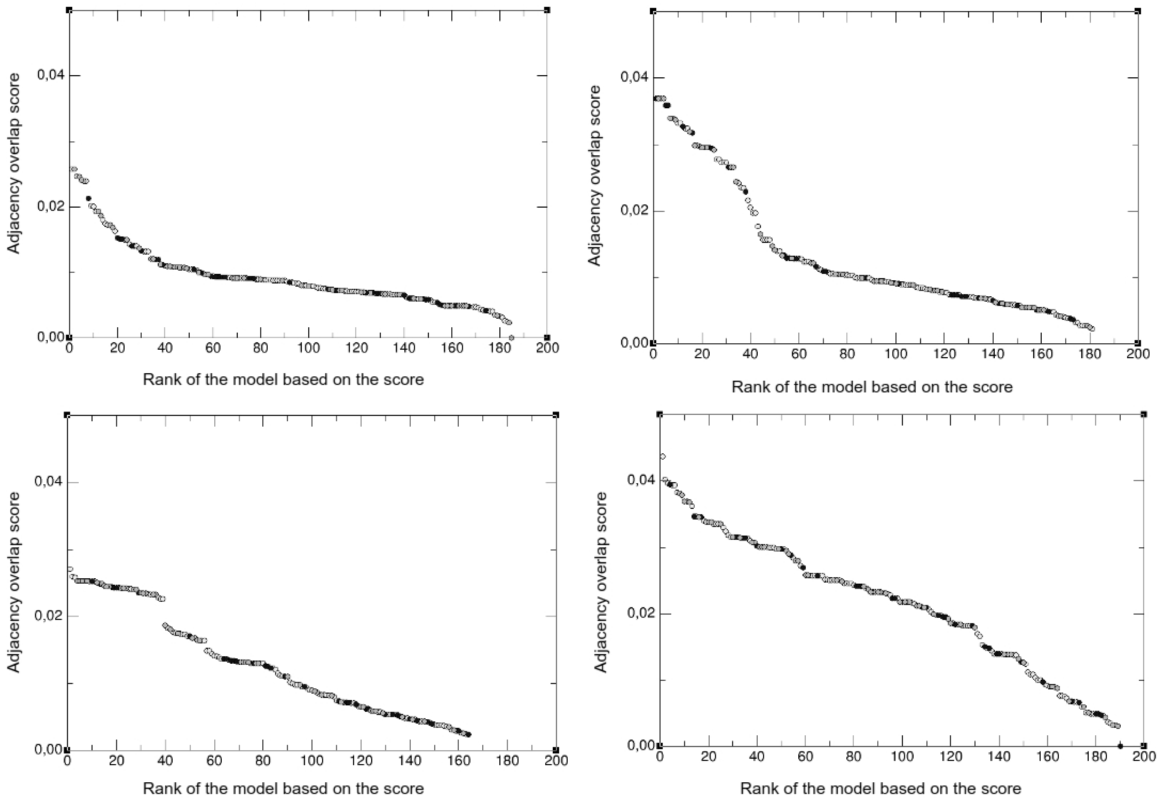


Fig. 4. Adjacency overlap results of T181 (**A**), T182 (**B**), T183 (**C**) and T184 (**D**). The colour corresponds to the ranking provided by the scorer group of the models from light grey meaning that models are ranked 5 or lower to dark grey where model is ranked 1.

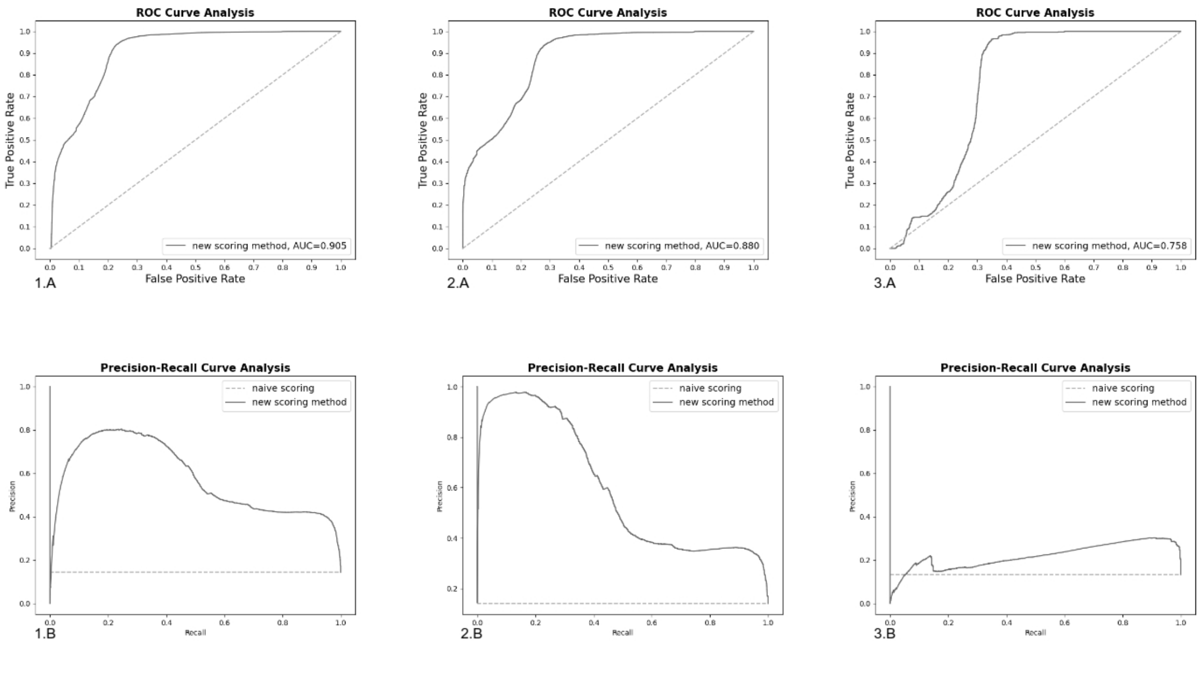


Fig. 5. ROC curves (A) and Precision-recall curves (B) of the full scoreset (1), the Eukaryotes set (2) and Viruses set (3). An Area Under Curve (AUC) is calculated for the ROC curves. The naive scorings of the precision recall curves correspond to the precision score if the models are randomly predicted to be good.

Session 3, Amphi C

Systems biology & metabolomics

Inference of an Integrative, Executable Network for Rheumatoid Arthritis Combining Data-Driven Machine Learning Approaches and a State-of-the-Art Mechanistic Disease Map

Quentin MIAGOUX[1], Vidisha SINGH[1], Dereck DE MÉZQUITA[1], Valerie CHAUDRU[1], Mohamed ELATI[2], Elisabeth PETIT-TEIXEIRA[1] and Anna NIARAKIS[1, 3]

[1] Université Paris-Saclay, Univ Evry, Laboratoire Européen de Recherche pour la Polyarthrite rhumatoïde-Genhotel, 91057 Evry, France

[2] CANTHER, University of Lille, CNRS UMR 1277, Inserm U9020, 59045 Lille, France

[3] Lifeware Group, Inria, Saclay-île de France, 91120 Palaiseau, France

Corresponding Author: anna.niaraki@univ-evry.fr

Paper Reference: **Miagoux *et al.* (2021) Inference of an Integrative, Executable Network for Rheumatoid Arthritis Combining Data-Driven Machine Learning Approaches and a State-of-the-Art Mechanistic Disease Map. Journal of Personalized Medicine. 2021; 11(8):785. https://doi.org/10.3390/jpm11080785**

1. Abstract

Rheumatoid arthritis (RA) is a multifactorial, complex autoimmune disease that involves various genetic, environmental, and epigenetic factors. Systems biology approaches provide the means to study complex diseases by integrating different layers of biological information. Combining multiple data types can help compensate for missing or conflicting information and limit the possibility of false positives. Our aim is to shed light on mechanisms governing the regulation of key transcription factors in RA and derive patient-specific models to gain more insights into the disease heterogeneity and the response to treatment.

In this work, we present a framework for integrating signalling and transcriptional regulation cascades with genomic mutations, combining data-driven approaches with prior knowledge in the form of an integrative RA-specific network. To do so, we use publicly available transcriptomic data of white blood cells from patients suffering from RA and the tool CoRegNet [1] to infer a co-regulatory network. Next, we develop an integration pairing method to couple the RA co-regulatory network with a state-of-the-art disease map for RA [2] to enrich the cooperativity network with upstream signalling regulators. Disease maps are comprehensive, knowledge-based representations of disease mechanisms, including disease-related molecular interactions supported by literature-based evidence [3]. Next, we project on the integrative RA network public genomic data and transcriptomic data from treated RA patients, highlighting key mutation carriers and differentially expressed genes associated with the response to anti-TNF treatment. Lastly, we study the system's dynamic behaviour using Boolean formalism to simulate subparts of the integrated network [4]. We perform in silico simulations, sensitivity analysis, and dose–response analyses to study the impact of other signalling cascades on the expression of the identified TFs, and steady-state analysis revealing combinations and conditions that can switch on or off the identified TFs, mimicking the effects of the treatment.

References

1. Remy Nicolle, Francois Radvanyi, and Mohamed Elati. CoRegNet: Reconstruction and Integrated Analysis of Co-Regulatory Networks. In *Bioinformatics* 31, 3066–3068, 2015.
2. Vidisha Singh, Giorgos D Kalliolias, Marek Ostaszewski, Maeva Veyssiere, Eleutherios Pilalis, Piotr Gawron, Alexander Mazein, Eric Bonnet, Elisabeth Petit-Teixeira, and Anna Niarakis. A. RA-Map: Building a State-of-the-Art Interactive Knowledge Base for Rheumatoid Arthritis. In *Database*, 2020.
3. Marek Ostaszewski, Stephan Gebel, Inna Kuperstein, Alexander Mazein, Andrei Zinovyev, Ugur Dogrusoz, Jan Hasenauer, Ronan M T Fleming, Nicolas Le Novère, Piotr Gawron, Thomas Ligon, Anna Niarakis, David Nickerson, Daniel Weindl, Rudi Balling, Emmanuel Barillot, Charles Auffray, and Reinhard Schneider, Community-driven roadmap for integrated disease maps. In *Briefings in Bioinformatics*, Pages 659–670, 2019.
4. Benjamin A. Hall and Anna Niarakis, Data integration in logic-based models of biological mechanisms. In *Current Opinion in Systems Biology,* 2021.

Calcium signaling is impaired in PTEN-deficient T cell acute lymphoblastic leukemia

Saran Pankaew[1,2], Delphine Potier[1], Clémence Grosjean[1], Mathis Nozais[1], Julie Quessada[1], Marie Loosveld[1,3], Élisabeth Remy[2*] & Dominique Payet-Bornet[1*]

[1] Aix Marseille Univ, CNRS, INSERM, CIML, 13288, Marseille, France

[2] Aix Marseille Univ, CNRS, I2M, 13288, Marseille, France

[3] APHM, Hôpital La Timone, Laboratoire d'Hématologie, 13385, Marseille, France

* Correspondence:
Élisabeth Remy (elisabeth.remy@univ-amu.fr)
Dominique Payet-Bornet (payet@ciml.univ-mrs.fr)

Paper Reference: **Pankaew *et al.* (2022) *Calcium signaling is impaired in PTEN-deficient T cell acute lymphoblastic leukemia,* Frontiers in Immunology, 2022, 13. https://doi.org/10.3389/fimmu.2022.797244**

PTEN (Phosphatase and TENsin homolog) is a phosphatase that antagonizes the action of Phosphatidylinositol 3 kinase (PI3K). It is the main negative regulator of PI3K/AKT signaling pathway and a well-known tumor suppressor involved in numerous types of cancer, including T-cell acute lymphoblastic leukemia (T-ALL) [1]. In humans, loss-of-function mutations of PTEN are correlated to mature T-ALL expressing a T-cell receptor (TCR) at their cell surface. In accordance with human T-ALL, the deletion of Pten gene in mouse thymocytes induces TCRαβ+ T-ALL development. In a previous study, we showed that, compared to physiological mature thymocytes, these Pten-deleted leukemic blasts display lower phosphorylated tyrosine level upon TCR activation [2]. This indicates that TCR signaling is impacted in Pten-deleted tumor cells. Herein, we explore the molecular mechanisms explaining this impact using conditional CD4-Cre x Pten$^{fl/fl}$ (Pten^{del}) mouse model. First, we performed multiplexed single-cell RNA sequencing (scRNAseq) of Pten^{del} and PTEN-proficient thymocytes. Bioinformatic analysis of our scRNAseq data show that pathological Pten^{del} thymocytes display, as expected, high Myc oncogene expression and an elevated PI3K/AKT pathway activity score. Moreover, the calcium pathway activity score of these Pten^{del} thymocytes is lower than in the physiological counterparts. We confirmed this result using ex vivo calcium flux assay and showed that upon TCR activation tumoral Pten^{del} blasts are unable to release Ca^{2+} into the cytoplasm. In order to understand such phenomena and the role of PTEN, we constructed a mathematical model centered on the mechanisms controlling the calcium flux, which integrates TCR signal strength and PTEN interactions. This qualitative model displays a dynamical behavior, coherent with the dynamics reported in the literature. Our model predicts that a loss of PTEN affects negatively IP3 (inositol 1,4,5-trisphosphate) receptors (IP3R). Hence, we analyzed IP3R protein expression by immunoblotting. In agreement with our model, our data show that IP3R protein level is reduced in PTEN-deficient tumor cells compared to physiological PTEN-proficient cells.

In conclusion, we showed that the calcium signaling pathway, known to be involved in thymocyte development and to be induced upon TCR engagement, is affected in PTEN-deficient TCRαβ+ T-ALL. Thanks to our mathematical modelling of calcium flux network, we propose that a functional interaction between PTEN and IP3R may explain the abrogation of calcium release upon TCR stimulation of PTEN-deficient leukemic cells.

References

1. Milella M, Falcone I, Conciatori F, Cesta Incani U, Del Curatolo A, Inzerilli N, Nuzzo CMA, Vaccaro V, Vari S, Cognetti F, et al. PTEN: Multiple Functions in Human Malignant Tumors. *Front Oncol*, **5**:24, 2015.
2. Gon S, Loosveld M, Crouzet T, Potier D, Bonnet M, Morin SO, Michel G, Vey N, Nunès JA, Malissen B, et al. Fit αβ T-cell receptor suppresses leukemogenesis of Pten-deficient thymocytes. *Haematologica*, 103:999–1007, 2018.

Comprehensive multi-omics integration of major depressive disorder

Amazigh MOKHTARI[1], El Cherif IBRAHIM[2], Arnaud GLOAGUEN[3], David COHEN[1], Margot DEROUIN[1], Cynthia MARIE-CLAIRE[4], Bruno ETAIN[4,5], Raoul BELZEAUX[5,6,7], Andrée DELAHAYE-DURIEZ[1,8,9] and Pierre-eric LUTZ[10,11]

[1] NeuroDiderot, Inserm U1141, Université de Paris, F-75019 Paris, France.
[2] Aix Marseille Université CNRS, Institut de Neurosciences de la Timone, F-13005 Marseille,France
[3] Centre National de Recherche en Génomique Humaine, Institut François Jacob, CEA, Université Paris-Saclay, Évry, France
[4] Université de Paris, INSERM UMR-S 1144, Optimisation thérapeutique en neuropsychopharmacologie, OTeN, F-75006 Paris, France
[5] Assistance Publique des Hôpitaux de Paris, GHU Lariboisière-Saint Louis-Fernand Widal, DMU Neurosciences, Département de psychiatrie et de Médecine Addictologique, F-75010 Paris, France
[6] Fondation FondaMental, F-94000 Créteil, France
[7] Assistance Publique Hôpitaux de Marseille, Pôle de psychiatrie, pédopsychiatrie et addictologie, F-13005 Marseille, France
[8] Assistance Publique des Hôpitaux de Paris, Unité de médecine génomique, Département BioPhaReS, Hôpital Jean Verdier, Hôpitaux Universitaires de Paris Seine Saint Denis, F-93140, Bondy, France
[9] Université Sorbonne Paris Nord, F-93000 Bobigny, France
[10] Centre National de la Recherche Scientifique, Université de Strasbourg, Fédération de Médecine Translationnelle de Strasbourg, Institut des Neurosciences Cellulaires et Intégratives UPR3212, F-67000 Strasbourg, France
[11] Douglas Mental Health University Institute, McGill University, QC H4H 1R3 Montréal, Canada

Corresponding author: andree.delahaye@inserm.fr , pierreeric.lutz@gmail.com

Abstract *Major depressive disorder (MDD) is a leading cause of disability worldwide with comorbidities such as an increased risk of suicide, drug dependence, and a reduced life expectancy. In addition to its heterogeneity, triggering and aggravating environmental factors complexify the identification of biomarkers and adequate therapeutics.*

Recently, the emergence of advanced approaches considering the biological complexity of multifactorial diseases broadened treatment targets in human cancer. Applied to psychiatric disorders, this may grant new insights on their pathophysiology and allow a more precise patient stratification.

In this study, we took advantage of transcriptomics (RNA-seq) and epigenomics (miRNA-seq and DNAm) datasets generated from a naturalistic cohort (n=65 MDD and n=81 controls) to build an integrative framework that includes late and intermediate integration methods. First, we applied sequential late integration with co-expression network and enrichment analyses. Second, we compared several joint dimensionality reduction learning methods to prioritize features that better discriminate between patients and controls. Additional clustering approaches were used on each set of selected variables and their performances were evaluated using an adjusted rand index (ARI).

The step-by-step integration revealed subnetworks associated with MDD, enriched for differentially expressed genes, methylated probes, and for targets of differentially expressed miRNAs. Furthermore, late integration identified features allowing a satisfactory clustering of cases and controls. Interestingly, both integration methods outputs (MDD associated co-expression networks and prioritized features that better discriminate MDD patients from controls) showed functional enrichments for common ontologies involving the immune response system. Applying advanced integration methods on large scale multi-omics psychiatric disorders cohorts could leverage better understating and treatment of MDD.

Keywords Multi-omics Integration; Major Depressive Disorder; DNA methylation; Transcriptomic; micro-RNA

1 Introduction

The Diagnostic and Statistical Manual of Mental Disorders (DSM-5) defines mood diseases as a collection of mental disorders characterized by recurrent and clinically significant changes in mood, energy, cognitive functions, sleep, or appetite. In the general population, Major Depressive Disorder

(MDD) is the most common syndrome among this category. Indeed, according to [1], MDD has a 12-month incidence of 10.4% and a lifetime prevalence of up to 20.6%. Due to their recurrence and commonly associated comorbidities, such as substance use disorders, anxiety disorders, and suicidal behaviors, mood disorders place a significant burden on patients and caregivers.

Therefore, individualized treatments to improve therapeutic results are crucial. Unfortunately, in psychiatry, the discovery of prognostic and diagnostic biomarkers is still in its early stages. Moreover, the multiple factors contributing to MDD, ranging from genetic predisposition to environmental influences, hampers the discovery of its underlying molecular mechanisms.

Lately, in the ever-growing Next Generation Sequencing (NGS) biological datasets, multi-omic data integration strategies have become necessary to identify the key processes that drive these interactions, and to prioritize putative biomarkers.

Multi-omics takes advantage of multidimensional datasets to their fullest extent, and has already improved the understanding of other multimodal diseases, such as cancer. Applied to psychiatry, it is expected to help refine diagnosis and patient stratification, thereby potentially improving therapeutic decisions.

Consequently, in the present work, we characterized and integrated multiple layers of transcriptomic and epigenetic (DNA methylation and micro-RNAs) data generated using peripheral blood samples from a cohort of individuals with MDD (N=80), or healthy controls (N=89). First, we analyzed each layer individually to extract single-omic signatures of MDD, and validated them by comparison with external datasets. Second, we characterized the modular and network organization of gene co-expression, and built a stepwise data integration framework by identifying gene modules whose co-expression associates with MDD, and show enrichment for differentially expressed or methylated genes. Finally, we used the Momix integration framework[2], which allows a more comprehensive view of omic interactions through matrix reduction methods, in order to extract features (individual mRNAs, micro-RNAs or DNA methylation probes) that most efficiently cluster MDD cases and controls. Collectively, our results provide support for the hypothesis that analyzing multiple molecular processes enables a better prediction of disease status than single-omic approaches, which has significant implications for the development of clinically useful multimodal biomarkers.

2 Material and methods summary

2.1 Data generation

The data generated for this manuscript were collected from peripheral blood of n=80 patients (meeting diagnostic criteria for MDD and suffering from a severe depressive episode at the time of blood sampling) and n= 89 healthy controls which are part of a naturalistic and longitudinal multicentric cohort registered in ClinicalTrials.gov with ID: NCT02209142.

2.1.1 RNA-seq Men, and women datasets were generated on different platforms, and then aligned and analyzed through a unified pipeline. Reads were trimmed with bbduk (bbmap package v38.73), and aligned to the GRCh38.p12 (hg38) reference genome using STAR (v2.5.3a) and comprehensive gene annotations from Gencode (v29). Gene quantification was then conducted using HTSeq (v0.11.2).

2.1.2 miRNA miRNA sequencing was performed for men and women at once on the same platform. Sequenced reads were trimmed using cutadapt (v1.18), following the QuickMIRseq pipeline. First, a small RNA and mRNA database were generated with sequences collected from the GRCh38.p12 (hg38) human reference genome. Then, a miRNA and hairpin database were generated with sequences collected from miRBase (v22-www.mirbase.org). Reads were aligned and quantified using bowtie (v1.2.2), with default parameters.

2.1.3 DNAm DNA methylation was assessed using Illumina Human MethylationEPIC BeadChip for men and women at once and on the same platform. Data analysis was performed using the R-package ChAMP (v2.16.2). Probes were filtered out when detection p-value above 0.01 in one or more samples, bead counts less than 3 in at least 5% of the samples, of non-CpG probes.

2.2 Differential analyses

Lowly expressed genes (mean raw expression < 10 reads) and miRNAs were first filtered out, and both miRNA-seq and RNA-seq data were analyzed for differential expression using the R package DESeq2. Differentially methylated probes (DMP) in EPIC arrays were identified with the ChAMP package. All data were adjusted for covariates, including: Age, BMI, batches, RNA integrity number (RIN), cellular composition and sex (except for RNAseq where differential expression analysis was conducted separately on each sex dataset). Functional enrichment analyses were performed using the Webgestalt web implementation of GSEA (n permutation = 10ç4) on GO, KEGG and Reactome databases.

2.3 Step-wise data integration

Weighted gene correlation network analysis (WGCNA) was performed in male and female separately to identify co-expression modules using gene expression levels, after first regressing out all covariates. WGCNA parameters were optimized to maximize overlaps with biological networks from the Reactome database, following recommendations by [3]. Gene modules were then tested for enrichments for men and women differentially expressed genes (DEG) in their respective datasets using GSEA, for targets of differentially expressed miRNA, and for DMPs, using a fisher overlapping test. Moreover, we used MAGMA to test for enrichment of gene modules in SNPs associated with psychiatric disorders. Summary statistics of Genome-Wide Association Studies (GWAS) of the Psychiatric Genomics Consortium were assigned to their nearest gene (within a 1 Mb window), and the enrichment of each module for those SNP-associated genes was then computed. Furthermore, associations of the eigengene of each module with case-control status, MDD and trauma scores derived from Childhood Trauma Questionaire (CTQ) were evaluated using Spearman correlation.

2.4 Advanced integration

Building on the Momix benchmark [2], we used six joint Dimension Reduction methods (jDR) methods (RGCCA, JIVE, MOFA, MOFA2, intNMF and SciKit-Fusion) to extract the common variance across the different omic modalities, their combinatorial and the severity of MDD, as reflected by the Hamilton Depression Rating Scale (HDRS) score. For each jDR method, a factor matrix F of dimensions $n \times k$ was derived, where n is the number of samples and $k = 10$ the number of factors extracted, as well as L weight matrices of dimension $p_j \times k$, $j = 1...L$, where p_j is the number of variables associated with the omic dataset j and L the number of omics modalities. Due to computational limitations, a pre-filtering step was applied to the DNAm data, where 10% of the most variable cgs were selected beforehand. A factor of interest F_i fixing the weight of the features was designated, from each of the F matrices, as the best correlating factor with the case control status. Then, only the top 10%-ranking features, according to their weights, were retained. Finally, omics dataset corresponding to the selected features were reconstructed and used to classify the datasets using two clustering methods, NEMO and SNF.

3 Results

3.1 Differential analysis results at the single omic level

As a first step, we derived lists of differentially expressed genes (DEG), micro-RNA (DEmiR) and differentially methylated probes (DMP) from each omics individually. Next, we tested for molecular signature differences between men and women using rank-rank hypergeometric overlap (RRHO2). Finally, results were compared to previously published studies and meta-analyses (see Fig.1 below).

3.1.1 mRNA differential expression analysis

A false discovery rate threshold (FDR < 0.05) was selected to identify DEGs between MDD patients and controls for each gender. 671 and 1034 genes were up-regulated and 273 and 790 were down-regulated in men and women, respectively (Fig1.a). Male and female DEGs were then compared to

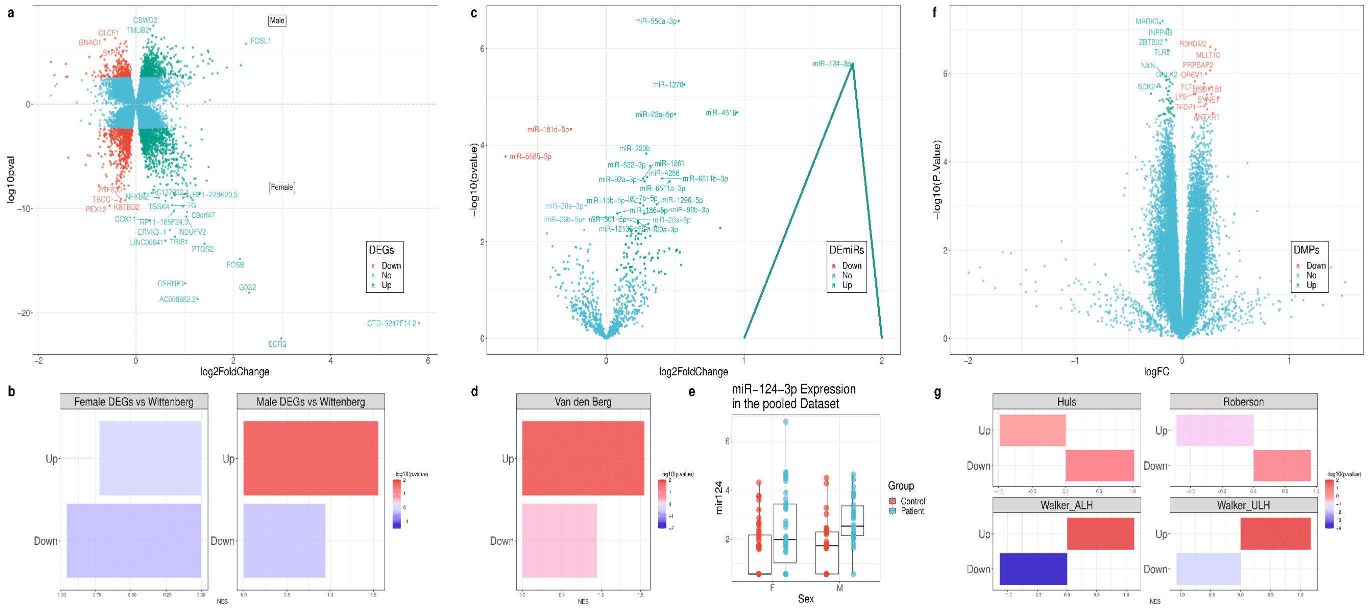


Fig. 1. Single-omics differential analyses results and enrichements: a. Male and female volcano plots of differentially expressed genes (threshold: P.adj<0.05). **b.** GSEA results of male and female DEG comparison with Wittenberg's quantitative review.**c.** Volcano plot of differentially expressed miRNA (threshold: P.adj<0.05). **d.** GSEA results of DEmiRNA comparison with van den Berg's results. **e.** Boxplots of miRNA-124-3p's normalized expression levels according to sex and case-control status. **f** Volcano plot of differentially methylated probes (threshold: P.adj<0.05). **g** GSEA results of DMPs comparison with Huls, Roberson and Walker results

Wittenberg's [4] quantitative review results of transcriptional differences in MDD case-control studies using peripheral blood samples (Fig.1.b). Male DEGs in our cohort showed significant enrichment for Wittenberg's up and down regulated gene lists (Up: NES=1.59 - P.adj=8.10E-04 ; Down: NES = -1.67 – P.adj=8.10E-04). While no significant enrichment was found for female DEGs.

Using GSEA, both male and female DEGs were tested for KEGG pathways enrichments. Women DEGs were significantly enriched for Aldosterone (NES=1.84, FDR=0.02) and Cortisol synthesis (NES=1.84, FDR=0.02) , as well as the Glutamatergic synapse (NES=1.86, FDR=0.02), while men's data showed enrichment for the immune response system through IL-17 signaling pathway (NES=1.75, FDR=7.58 10E-03) and neurodegenerative diseases (including Huntington, Parkinson and Alzheimer disease - NES=1.86, FDR=1.71 10E-03).

3.1.2 miRNA differential expression analysis Out of the 735 miRNAs that met quality control, 14 were differentially expressed in the pooled cohort (male and female) among which 12 were up-regulated and 2 were down-regulated (threshold: -log10(P.value) > 3 - Fig1.c). Comparison of our results with those previously reported in relation to MDD [5] found significant enrichment for up-regulated miRNA (Fig.1d). Noteworthy, miR-124-3p, a miRNA frequently associated with MDD, was the only one significantly up-regulated in both men and women (Fig.1e).

3.1.3 DNAm differential methylation analysis Of the 850k CG methylation sites investigated (Fig.1f), only a scarce number met a threshold FDR < 0.1 (pooled cohort: 70, male: none , female: 155). Similar proportions of hypo- or hyper- methylation events were observed in women, while hypo-DMPs were more frequent in men. Comparison of DMP with other studies showed significant enrichment in two out of the four dataset tested (Fig.1g).

3.2 Comparison of MDD-related differential expression across men and women

A threshold-free approach (rank-rank hypergeometric overlap - RRHO2) was used to test for similar and opposite expression and methylation profiles in male and female (Fig.2). Throughout the analysis, this method identified statistically significant overlaps for up regulated miRNAs, mRNA and

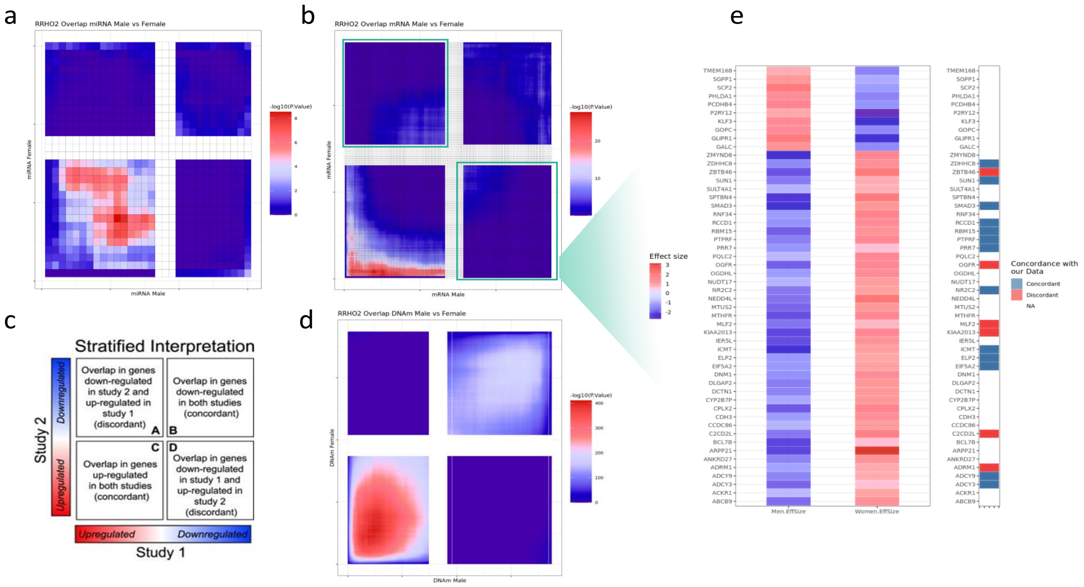


Fig. 2. Rank-rank hypergeometric overlap results: a. miRNA male vs female comparaison; **b.** mRNA male vs female comparaison; **c.** interpretation of the RRHO results; **d.** DNAm male vs female comparaison; **e.** Overlap of male/female discordant gene with Seney's sex specific depression signature.

CpGs that were modified in similar directions in both men and women (bottom-left and upper-right frame Fig.2.a-c). In parallel, we also identified genes showing opposite expression changes across men and women as a function of MDD (upper-left and bootom-right frame). Then, we compared these sex-discordant genes identified in the blood with those previously identified in the brain by Seney's [6] opposite molecular signatures of depression in men and women. (Fig.2d). Among these, 21 were overlapping with our results, of which 13 showed concordant profile alterations with the sex molecular signature of MDD.

3.3 Early and step-wise integration identified MDD-associated modules that are enriched for immune response system KEGG term

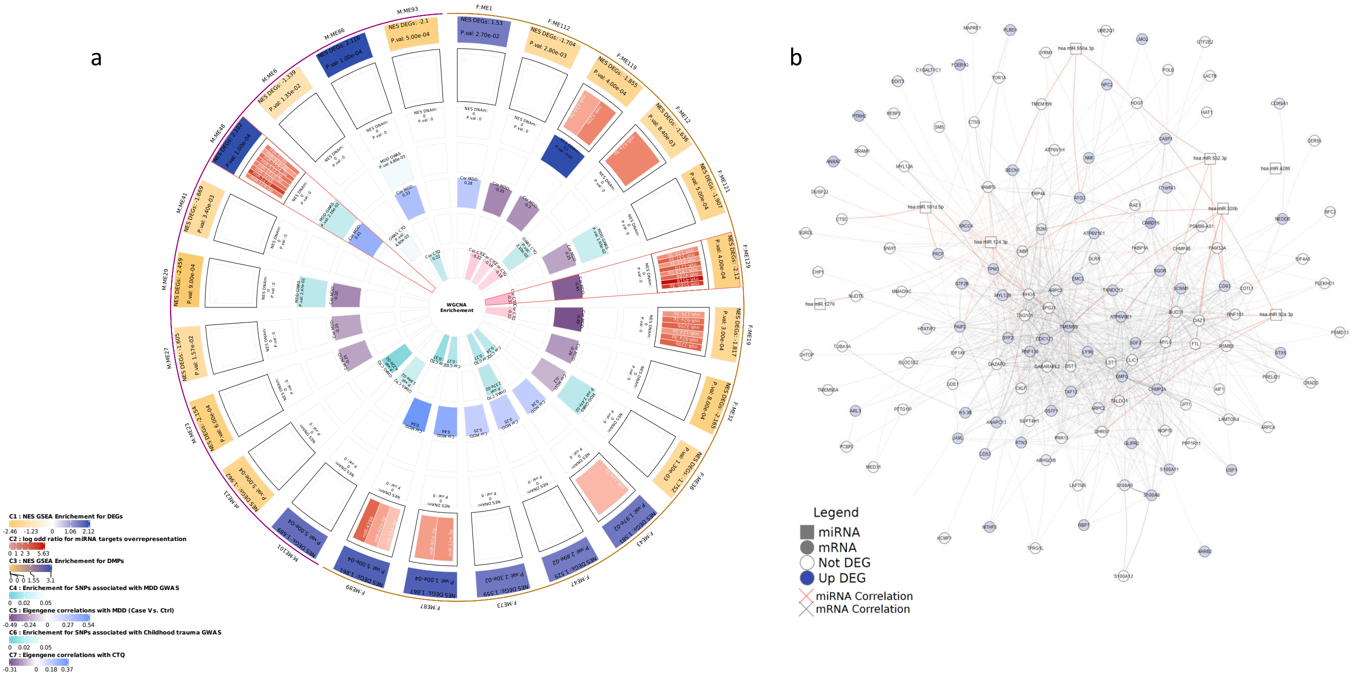


Fig. 3. Step-wise integration results: a. Circleplot of male and female WGCNA top modules enrichements; **b.** Graphical representaion of the M:ME48 module

To reduce data dimensionality, and conduct a multi-omics analysis, we inferred from the RNA-seq data highly co-expressed gene networks using WGCNA, applied in male and female separately. The resulting modules were then tested for enrichment for our two epigenetic datasets and clinical data (score of depression and childhood trauma). Furthermore, we determined which modules harbored significant association for genetic variation implicated in psychiatric disease, using external GWAS data. We identified multiple modules that are highly significantly enriched for DEGs and for validated targets of differentially expressed miRNAs (Fig3.a). Several modules exhibited significant association with both MDD and bipolar disorder GWAS results.

Among most noticeable modules, M:ME48 (181 genes, men module) and F:ME129 (66 genes, female module) exhibited enrichment for MDD up-regulated DEGs (NES = 2.10, P.adj = 0.0001) and MDD down-regulated DEGs (NES = -2.11 , P.adj = 0.0004), respectively, while both modules were also enriched for validated targets of miR-124-3p (M:ME48 :P.adj = 4.5 10-27 , F:ME129P.adj = 2.31 10-13). Moreover, M:ME48 showed enrichment for genes associated with SNPs identified by MDD GWAS studies (p-val = 0.002) while F:ME129 was enriched for SNPs involved in Bipolar disorder type 2. Interestingly, the module's eigengenes showed opposite significant correlations to case control status. (M:ME48: R2 = 0.4 , P.adj = 0.0007 ; F:ME129 : R2 = -0.45 , P.adj = 1.25 10-06). These two distinct modules exhibited a low overlap (Jaccard index = 0.1).

A graphical representation of the M:ME48 module (Fig.3b) shows hub genes that are differentially expressed, as well as validated targets of differentially expressed microRNAs. Moreover, M:ME48 showed enrichment for biological processes ontologies involving the immune response system.

3.4 Advanced integration with semi-supervised clustering prioritize mRNAs and DNA methylation CpGs as more efficient biomarkers of MDD in women

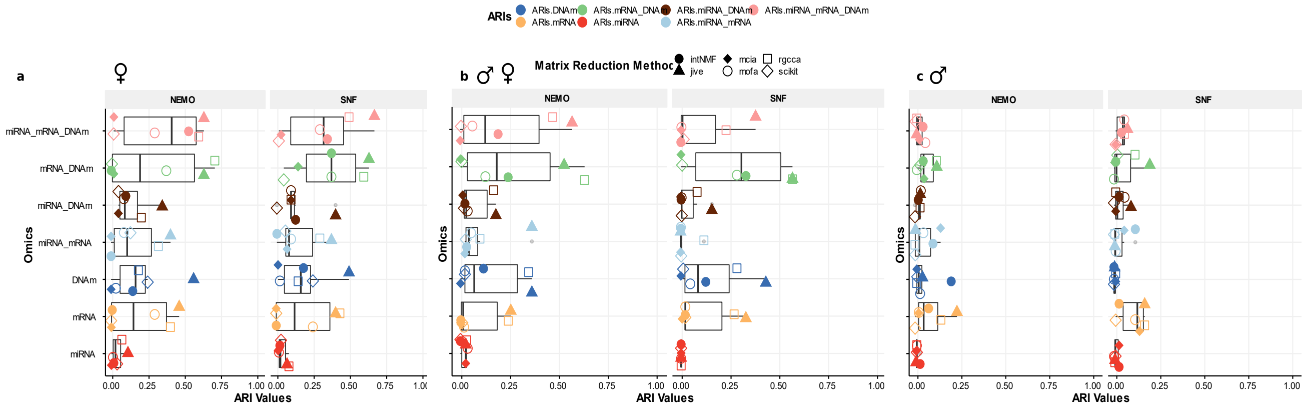


Fig. 4. MDD case-control clustering results. a-c Boxplots of ARIs values from NEMO and SNF clustering according to the used omics and matrix reduction methods in the woman , man and datasets.

Here, a semi-supervised framework was designed to select the best set of omic features in order to discriminate between MDD patients and controls. First, using the six joint dimension reduction (jDR) methods of the Momix framework [2],we identified features (miRNA, mRNA and CpG probes) that contribute the most to covariance in the three omics, when combined with measurements of depression severity (HDRS score). We compared the use of only one, two or three molecular layers as input data for each of the 6 jDR methods (in total 42 different modalities). Second, the top 10% highest contributing features in each modality were used to cluster the samples using two methods: NEMO and SNF.

The results of each omic combination is displayed in Fig.4. Regarding the methods, JIVE and RGCCA systematically outperformed the other approaches when applied to our dataset. In the combined cohort (Fig.4.b), the combination of mRNA and DNAm features identified by RGCCA yielded the highest ARI score (ARI = 0.63). Our framework performed poorly on the man dataset, independently of the used omic or methods Fig.4.c, where the highest ARI value was equal to 0.29 obtained from the combined three omics using intNMF.

The combined omics features derived from the pooled datasets, using JIVE, (surrounded in Fig.4.b) were assessed for stability using a bootstrapping approach (n=500). Subsequently, the stability of features were assessed using bootstrapping. Overall, mRNA and DNAm features showed satisfying stability results, while miRNA lacked of weighting consistency. Moreover, similarly to the DEGs, the top 1% of the mRNA features exhibited, as best GO terms, enrichments involved in the immune system response.

4 Discussion and conclusion

We describe here a new vertical multi-omics dataset generated from MDD patients and controls. Standard single-omic analyses replicated previously reported biological signals associated in peripheral blood with MDD, providing external validation of this new MDD dataset. To take advantage of the three molecular levels measured from the same blood samples, we used both step-by-step and advanced multi-omic integration approaches

For each single molecular layer, differential analyses showed significant sex differences, as well as different enrichments. In men, transcriptomic analysis, showed concordant enrichments with previously described pathways, mainly involving the immune system response [7] as well as enrichments for neurodegenerative diseases[8]. In women, data revealed enrichments involving cortisol and aldosterone synthesis [9], as well as the glutamatergic synapse [10]. Despite discrepancies among man and woman enrichments , up-regulated genes in the two datasets demonstrated a significant rank-rank hypergeometric overlap. Regarding sex-discordant genes, they overlapped with a sex-specific MDD signature. This sex-dependent variation could be explained by a difference in the molecular signature of depression according to gender, as described in the literature [11].

Regarding miRNAs, miR-124-3p was previously described as a potential biomarker of psychiatric disorders [12] and was systematically up-regulated significantly in the distinct and pooled male and female datasets. In particular , the association of elevated miR-124 -3p expression in the blood with the stress response and alterations of neural plasticity have been previously described in patients with MDD [12]. Our results strengthen the rationale for using this miR as a potential biomarker of depression.

A limited number of DMPs were found in MDD patients compared to controls in our study. This could be explained by the relatively small size of our cohort, which was unable to capture the subtle variations of DNA methylation in peripheral tissues of depressive patients. Nevertheless, these DMPs were significantly enriched for previously reported DMPs in MDD patients [13]) and interestingly for olfactory transduction. This is in line with the difference in overall olfactory functioning between depressed individuals and non-depressed controls [14].

In the second part of this study, we first integrated the results of the aforementioned analyses through a step-by-step integrative approach. The integration revealed two gene modules from the male and female data that presented interesting functional enrichments involving mainly the immune system response. The male module M:ME48 was found enriched in differentially expressed gene targets of miR 124-3p, as well as in genes associated with SNPs implicated in the risk for MDD. This gene network therefore appears of strongest interest as a molecular signature of MDD.

In the last part of our analysis, we investigated the potential contribution of several omics modalities in the clustering of our datasets. The joint dimension reduction approach allowed the selection of features allowing for the classification of cases and controls. Despite some satisfying ARI results from the pooled cohort, the clustering performed poorly when considering the man cohort only, in comparison to using the woman one only. Furthermore, in both pooled and female datasets, the selected features belonged to the transcriptomic and DNA methylation layers, excluding the miRNome. Results therefore show substantial contribution from the use of these two modalities over single omic based analyses.

In conclusion, our study is the first vertical multi-omic integration of major depressive disorder. Our analyses show a sex dependent biological signal of MDD that is important to consider when seeking to identify MDD biomarkers. Finally, most importantly, we also demonstrate an information gain in the use of multi-omics for classification of MDD over single omics approaches. These results could serve as a baseline for the development of classifiers helping the stratification of MDD patients.

References

[1] Deborah S. Hasin, Aaron L. Sarvet, Jacquelyn L. Meyers, Tulshi D. Saha, W. June Ruan, Malka Stohl, and Bridget F. Grant. Epidemiology of adult DSM-5 major depressive disorder and its specifiers in the united states. *JAMA Psychiatry*, 75(4):336, April 2018.

[2] Laura Cantini, Pooya Zakeri, Celine Hernandez, Aurelien Naldi, Denis Thieffry, Elisabeth Remy, and Anaïs Baudot. Benchmarking joint multi-omics dimensionality reduction approaches for the study of cancer. *Nature Communications*, 12(1), January 2021.

[3] Tooba Abbassi-Daloii, Hermien E. Kan, Vered Raz, and P.A.C. 't Hoen. Recommendations for the analysis of gene expression data to identify intrinsic differences between similar tissues. *Genomics*, 112(5):3157–3165, September 2020.

[4] Gayle M. Wittenberg, Jon Greene, Petra E. Vértes, Wayne C. Drevets, and Edward T. Bullmore. Major depressive disorder is associated with differential expression of innate immune and neutrophil-related gene networks in peripheral blood: A quantitative review of whole-genome transcriptional data from case-control studies. *Biological Psychiatry*, 88(8):625–637, October 2020.

[5] M.M.J. van den Berg, J. Krauskopf, J.G. Ramaekers, J.C.S. Kleinjans, J. Prickaerts, and J.J. Briedé. Circulating microRNAs as potential biomarkers for psychiatric and neurodegenerative disorders. *Progress in Neurobiology*, 185:101732, February 2020.

[6] Marianne L. Seney, Zhiguang Huo, Kelly Cahill, Leon French, Rachel Puralewski, Joyce Zhang, Ryan W. Logan, George Tseng, David A. Lewis, and Etienne Sibille. Opposite molecular signatures of depression in men and women. *Biological Psychiatry*, 84(1):18–27, July 2018.

[7] Eléonore Beurel, Marisa Toups, and Charles B. Nemeroff. The bidirectional relationship of depression and inflammation: Double trouble. *Neuron*, 107(2):234–256, July 2020.

[8] Ciaran Galts, Luis Bettio, David C. Jewett, Charles C. Yang, Patricia S. Brocardo, Ana Lucia S. Rodrigues, Jonathan S. Thacker, and Joana Gil-Mohapel. Depression in neurodegenerative diseases: Common mechanisms and current treatment options. *Neuroscience & Biobehavioral Reviews*, 102:56–84, July 2019.

[9] Jan Nowacki, Katja Wingenfeld, Michael Kaczmarczyk, Woo Ri Chae, Paula Salchow, Ikram Abu-Tir, Dominique Piber, Julian Hellmann-Regen, and Christian Otte. Steroid hormone secretion after stimulation of mineralocorticoid and NMDA receptors and cardiovascular risk in patients with depression. *Translational Psychiatry*, 10(1), April 2020.

[10] Sho Moriguchi, Akihiro Takamiya, Yoshihiro Noda, Nobuyuki Horita, Masataka Wada, Sakiko Tsugawa, Eric Plitman, Yasunori Sano, Ryosuke Tarumi, Muhammad ElSalhy, Nariko Katayama, Kamiyu Ogyu, Takahiro Miyazaki, Taishiro Kishimoto, Ariel Graff-Guerrero, Jeffrey H. Meyer, Daniel M. Blumberger, Zafiris J. Daskalakis, Masaru Mimura, and Shinichiro Nakajima. Glutamatergic neurometabolite levels in major depressive disorder: a systematic review and meta-analysis of proton magnetic resonance spectroscopy studies. *Molecular Psychiatry*, 24(7):952–964, October 2018.

[11] George M. Slavich and Julia Sacher. Stress, sex hormones, inflammation, and major depressive disorder: Extending social signal transduction theory of depression to account for sex differences in mood disorders. *Psychopharmacology*, 236(10):3063–3079, July 2019.

[12] Bhaskar Roy, Michael Dunbar, Richard C Shelton, and Yogesh Dwivedi. Identification of MicroRNA-124-3p as a putative epigenetic signature of major depressive disorder. *Neuropsychopharmacology*, 42(4):864–875, August 2016.

[13] Rosie May Walker, Andrea Nikie Christoforou, Daniel L. McCartney, Stewart W. Morris, Nicholas A. Kennedy, Peter Morten, Susan Maguire Anderson, Helen Scott Torrance, Alix Macdonald, Jessika Elizabeth Sussmann, Heather Clare Whalley, Douglas H. R. Blackwood, Andrew Mark McIntosh, David John Porteous, and Kathryn Louise Evans. DNA methylation in a scottish family multiply affected by bipolar disorder and major depressive disorder. *Clinical Epigenetics*, 8(1), January 2016.

[14] Hannah Taalman, Caroline Wallace, and Roumen Milev. Olfactory functioning and depression: A systematic review. *Frontiers in Psychiatry*, 8, September 2017.

Session 4, Amphi A

Statistics, machine learning, AI & image analysis II

Computing disease progression scores using multimodal variational autoencoders trained with neuroimaging and microRNA data

Virgilio KMETZSCH[1,2], Emmanuelle BECKER[3], Dario SARACINO[1,2], Vincent ANQUETIL[2], Daisy RINALDI[2], Agnès CAMUZAT[2], Thomas GAREAU[2], Isabelle LE BER[2], Olivier COLLIOT[2,1] and The PREV-DEMALS study group

[1] Inria, Aramis project-team, F-75013, Paris, France

[2] Sorbonne Université, Paris Brain Institute - Institut du Cerveau - ICM, Inserm U1127, CNRS UMR 7225, AP-HP - Hôpital Pitié-Salpêtrière, Paris, France

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

Corresponding author: olivier.colliot@sorbonne-universite.fr

Reference paper:* Kmetzsch *et al.* (2022) A multimodal variational autoencoder for estimating progression scores from imaging and microRNA data in rare neurodegenerative diseases. *SPIE Medical Imaging 2022. https://hal.archives-ouvertes.fr/hal-03576117

1 Introduction

Frontotemporal dementia (FTD) and amyotrophic lateral sclerosis (ALS) are rare neurodegenerative disorders, frequently hereditary, and have no effective treatment. Their most prevalent genetic cause are mutations in the *C9orf72* gene. Since no individual biomarker can accurately measure progression in these disorders, it is essential to incorporate data from many modalities when evaluating novel therapy strategies. Particularly, neuroimaging and microRNA data have been demonstrated to be useful in tracking FTD and ALS progression [1], [2]. As these conditions are rare, algorithms that fuse multimodal data from small samples must be developed.

2 Methods

We present a technique for generating a disease progression score (DPS) from cross-sectional multimodal data, based on variational autoencoders (VAE) [3]. Our hypothesis is that VAEs trained with neuroimaging and microRNA data from small samples are able to estimate a meaningful latent space, where the relative positions of latent codes represent the distance traveled along the underlying FTD/ALS pathophysiological pathway. A DPS can thus be inferred by orthogonal projections into a trajectory in the latent space. Models were tested with multimodal data from the PREV-DEMALS study (14 patients, 40 presymptomatic mutation carriers, and 37 healthy controls), consisting of expression levels of 589 microRNAs and grey matter volumes of 87 brain regions. Since the DPS has no ground truth, we used the inferred scores to do pairwise classifications, and computed the areas under the ROC curves as proxy performance metrics.

3 Results

After a 5-fold cross-validation using all available features, presymptomatic subjects and patients were classified with an average area under the ROC curve of 0.83. Moreover, when we reduced the dimension of the microRNA data to 68 with a discovery set, the average ROC AUC was 0.94.

4 Conclusion

Although long-term follow-up data will be needed to assess the accuracy of the inferred DPS, our results indicate that the presented framework can leverage cross-sectional multimodal datasets with small sample sizes in order to measure disease progression.

References

[1] Virgilio Kmetzsch, Vincent Anquetil, Dario Saracino, et al. Plasma microRNA signature in presymptomatic and symptomatic subjects with C9orf72-associated frontotemporal dementia and amyotrophic lateral sclerosis. *Journal of Neurology, Neurosurgery & Psychiatry*, 92(5):485–493, May 2021.

[2] Anne Bertrand, Junhao Wen, Daisy Rinaldi, et al. Early Cognitive, Structural, and Microstructural Changes in Presymptomatic C9orf72 Carriers Younger Than 40 Years. *JAMA neurology*, 75(2):236–245, 2018.

[3] Diederik P. Kingma and Max Welling. Auto-Encoding Variational Bayes. *arXiv:1312.6114 [cs, stat]*, May 2014. arXiv: 1312.6114.

LowKi: Efficient estimation of kinship and fraternity coefficients from low-depth sequencing data

Anthony F. Herzig[1], Marina Ciullo[2,3], FranceGenRef Consortium[4], Anne-Louise Leutenegger[5], and Hervé Perdry[2]

[1] Inserm, Univ Brest, EFS, UMR 1078, GGB, F-29200 Brest, France
[2] Institute of Genetics and Biophysics A. Buzzati-Traverso - CNR, Naples, Italy
[3] IRCCS Neuromed, Pozzilli, Isernia, Italy
[4] LABEX GENMED, Centre National de Recherche en Génomique Humaine, Evry, Paris
[5] Inserm, Université Paris Cité, UMR 1141, NeuroDiderot, F-75019 Paris, France
[6] CESP Inserm U1018, Université Paris-Saclay, Villejuif, France

Corresponding Author: anthony.herzig@inserm.fr

Abstract *We present LowKi (**Low**-depth **Ki**nship), an R-package for directly and rapidly estimating kinship and fraternity coefficients from low-depth sequencing data in variant-call format (https://github.com/genostats/LowKi). Estimating such coefficients is an important step for many study designs and analyses of genetic datasets and a large variety of methods have been proposed for their estimation from genotype data. When dealing with sequencing data, one can call the genotypes, and then apply these methods to the resulting dataset. However, individual genotypes cannot be confidently called from low-depth sequencing data. Ignoring this uncertainty leads to biased estimates. A small number of methods have been developed to estimate kinship and fraternity specifically with low-depth sequencing data by basing calculations on genotype likelihoods; this is precisely the aim of LowKi. What differentiates LowKi is its efficiency; requiring far less runtime than other software and does not require the use of intermediate software. This is made possible by firstly using method-of-moment estimates that are quick to calculate and secondly by using an innovative regression-based bias adjustment procedure allowing LowKi to equal competing software in terms of accuracy.*

Keywords Low-depth sequencing, Kinship, Relatedness, Fraternity, Next-generation sequencing, genotype uncertainty

1. Introduction

Correctly estimating relatedness coefficients is important for a wide range of applications in plant, animal, or human genetics. Coefficients such as kinship and fraternity describe the extent of shared recent origin between genomes. The most frequent uses of such coefficients are based on the genetic relationship matrix (GRM) of a sample of individuals, which is the symmetric matrix of such coefficients. Its Eigen-decomposition provides principal components which are commonly used to infer population structure [1] and for identifying internal clusters of similar individuals [2] or for controlling for inflation in association testing arising from population stratifications. GRMs themselves can be used as variance component for linear mixed modelling for correcting for stratification [3] or for heritability estimation [4–7]. Finally, estimating such coefficients allows for the identification of closely related individuals in the sample [8]; and in some cases could also help to identify contaminated samples [9].

Whole-genome sequencing (WGS) costs have fallen drastically in recent decades but now the price has begun to plateau [10,11]. Sequencing at a lower depth (number of times each genetic base is analyzed, denoted herein as $N\times$ for a sequencing depth of N) can reduce the cost, and hence a natural trade-off arises between the number of individuals to be sequenced and sequencing depth. When wishing to characterize a population or examine subtle differences in allele-frequencies it may well be preferable to have a very large sample with less-than-perfect data quality than a small sample with very precise data; hence the appeal of low-depth whole-genome sequencing (LDWGS).

To respond to the trend of LDWGS studies, we have developed LowKi [12] for estimating kinship and fraternity coefficients from such data. Kinship describes the probability of two randomly drawn alleles from a pair of individuals sharing a recent common ancestor. Fraternity describes the proportion of the genome where a pair of individuals share two pairs of alleles both descended from recent common ancestors. For example, a parent-offspring pair have a kinship coefficient of 0.25 and a fraternity coefficient of 0 whilst a sibling pair also have a kinship coefficient of 0.25 but a fraternity coefficient of 0.25. LowKi uses moment-estimates, is far quicker than existing software that provide the same service, and is competitive in terms of accuracy.

2. Methods

2.1. Moment-estimators from LDWGS data

LowKi's moment-estimators are similar to those used for genotype data but are based on genotype-likelihoods as are often provided from standard software such as GATK HaplotypeCaller [13]. Typically, the (twice) kinship matrix K and the fraternity matrix D can be estimated as follows:

$$K_{ii'} = \frac{1}{M}\sum_{j=1}^{M} X_A^{ij} \times X_A^{i'j} \quad \& \quad D_{ii'} = \frac{1}{M}\sum_{j=1}^{M} X_D^{ij} \times X_D^{i'j}$$

$K_{ii'}$ is twice the estimated kinship coefficient of individuals i and i'; and $D_{ii'}$ is their estimate of fraternity. The estimators are averages of point-wise estimated from M biallelic genetic variants, computed from the correlation of centered and reduced additive and non-additive genotype components X_A^{ij} and X_D^{ij} (for individual i and variant j) that are defined as follows:

$$\begin{aligned} X_A^{ij} &= \alpha_0^j\, 1_{\{G^{ij}=0\}} + \alpha_1^j\, 1_{\{G^{ij}=1\}} + \alpha_2^j\, 1_{\{G^{ij}=2\}} \\ X_D^{ij} &= \delta_0^j\, 1_{\{G^{ij}=0\}} + \delta_1^j\, 1_{\{G^{ij}=1\}} + \delta_2^j\, 1_{\{G^{ij}=2\}} \end{aligned}$$

where

$$\alpha_k^j = \frac{k - 2q_j}{\sqrt{2p_j q_j}}, (k = 0, 1, 2) \; and \;\; \delta_0^j = \frac{q_j}{p_j}, \delta_1^j = -1, \delta_2^j = \frac{p_j}{q_j},$$

q_j being the minor allele frequency of variant j and $p_j = 1 - q_j$. To form similar estimates for LDWGS data, we form different additive and non-additive components (that remain orthogonal by construction) that are not based on allele-frequencies but use sample means of genotype-frequencies estimated using genotype-likelihood data. The approach is similar to Vitezica et al. [14].

The additive and non-additive components take the respective forms,

$$\begin{aligned} \tilde{X}_A^{ij} &= \tilde{\alpha}_0^j\, P_{AA}^{ij} + \tilde{\alpha}_1^j\, P_{Aa}^{ij} + \tilde{\alpha}_2^j\, P_{aa}^{ij} \\ \tilde{X}_D^{ij} &= \tilde{\delta}_0^j\, P_{AA}^{ij} + \tilde{\delta}_1^j\, P_{Aa}^{ij} + \tilde{\delta}_2^j\, P_{aa}^{ij} \end{aligned}$$

where P_G^{ij} is the estimated probability of individual i having genotype G (three possibilities exist, 'AA', 'Aa', and 'aa' for a biallelic variant with wild-type allele 'A' and alternative allele 'a') for variant j. Here,

$$\left(\tilde{\alpha}_0^j, \tilde{\alpha}_1^j, \tilde{\alpha}_2^j\right) = \left(\bar{P}_{Aa}^j + 4\bar{P}_{aa}^j \bar{P}_{AA}^j - {\bar{P}_{Aa}^j}^2\right)^{-\frac{1}{2}} \times \left(-\bar{P}_{Aa}^j - 2\bar{P}_{aa}^j, 1 - \bar{P}_{Aa}^j - 2\bar{P}_{aa}^j, 2 - \bar{P}_{Aa}^j - 2\bar{P}_{aa}^j\right)$$

and

$$\left(\tilde{\delta}_0^j, \tilde{\delta}_1^j, \tilde{\delta}_2^j\right) = \left(\bar{P}_{aa}^j + 4\frac{\bar{P}_{aa}^j \bar{P}_{AA}^j}{\bar{P}_{Aa}^j} + \bar{P}_{AA}^j\right)^{-\frac{1}{2}} \times \left(\sqrt{\frac{\bar{P}_{aa}^j}{\bar{P}_{AA}^j}}, -2\sqrt{\frac{\bar{P}_{aa}^j \bar{P}_{AA}^j}{{\bar{P}_{Aa}^j}^2}}, \sqrt{\frac{\bar{P}_{AA}^j}{\bar{P}_{aa}^j}}\right).$$

Where $\bar{P}_G^j$ is the sample mean of values P_G^{ij} for genotype $G \in \{AA.Aa.aa\}$.

LowKi's moment estimators are finally given as:

$$\widetilde{K}_{ii'} = \frac{1}{M}\sum_{j=1}^{M} \tilde{X}_A^{ij} \times \tilde{X}_A^{i'j} \text{ and } \widetilde{D}_{ii'} = \frac{1}{M}\sum_{j=1}^{M} \tilde{X}_D^{ij} \times \tilde{X}_D^{i'j}.$$

2.2. Regression-cased bias adjustment

The moment-estimators of LowKi are subject to a multiplicative bias, this was readily observed through simulation studies and tests on real data. To account for this, we used an innovative regression-based solution. We regressed point-wise estimators (of the form $\tilde{X}_A^{ij} \times \tilde{X}_A^{i'j}$) against a measure of the genotype-uncertainty at variant in question for the two individuals. The key idea is that as the uncertainty goes to zero, the point-wise estimator should become unbiased - hence the intercept from the regression model provided an unbiased estimator. To prevent extensive calculations, we demonstrated that such regression need only be performed for a small sample of pairs in order to give an accurate computation of the multiplicative bias that the initial moment-estimators suffer from - hence we could efficiently readjust all pairs to provide accurate relatedness matrices quickly.

3. Results

To demonstrate LowKi, we present results from a small simulated dataset, corresponding to a fictitious small isolated population. This dataset is included in the R-package and an online vignette permits users to recreate the analyses given here. To build this simulation dataset, five generations of a panmictic population of 1000 individuals per generation were generated by gene dropping; the founders haplotypes were built as mosaics of European 1000 Genome Project [15] haplotypes using the R-package Mozza (https://github.com/genostats/Mozza). The dataset consists in a 68 Mb vcf file containing low-depth data (mean depth 1× to 3×) for 200 individuals taken from the last two simulated generations, with data in 200,000 autosomal SNPs only – to keep the file reasonably small. Two small 'rds' files are also included, corresponding to the 'true' kinship and fraternity matrices, as computed from the length of genome shared IBD by the simulated individuals.

To compute the kinship and fraternity matrices in LowKi, we simply have to run the respective R command:

```
K.low <- lowKi(vcf.file) ;  D.low <- lowKi(vcf.file,fraternity=TRUE)
```

In a first instance, we can compare the estimates of LowKi against a classic GRM estimated using hard-called genotypes (where the calling of LDWGS is forced by taking the genotype of maximal probability). LowKi is clearly more accurate than GRMs based on hard-calls (Fig. 1 & 2).

Using similar simulation datasets, we could show that LowKi could produce good estimates of Kinship down to an average depth of 0.1× and good estimates of fraternity down to an average depth of 1×. We determined that the current leading software for estimating kinship and fraternity from LDWGS was NGSRelateV2 [16]; this software had similar limitations in performance to LowKi (Fig. 3 & 4).

We had also tested SEEKIN [17] which only estimated kinship and requires a preliminary hidden-Markov model software BEAGLE [18] - which give it a very high de facto runtime. NGSRelateV2 was very accurate but also far slower than LowKi, on a simulated dataset of 1,444 individuals we noted that LowKi required 45 minutes, while SEEKIN required 34 hours and NGSRelateV2 required 13 days of runtime.

4. Conclusion

LowKi provides accurate estimation of kinship and fraternity directly from LDWGS data. It is based on fast-to-compute moment estimators, and on an innovative regression based method which removes the bias from these estimators. The main interest of LowKi will be in analysing large samples, for which it greatly outperforms its competitors in terms of run-time.

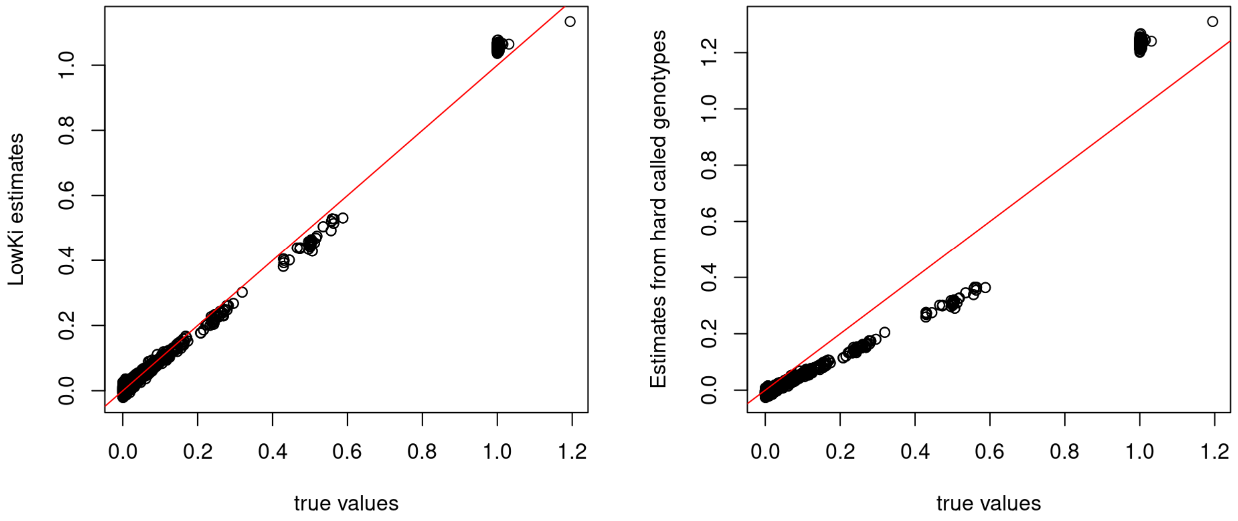


Fig 1. Estimating the kinship matrix, LowKi against a GRM based on hard-called genotypes, 'true values' on the x-axis indicate the simulated IBD-sharing between pairs of individuals.

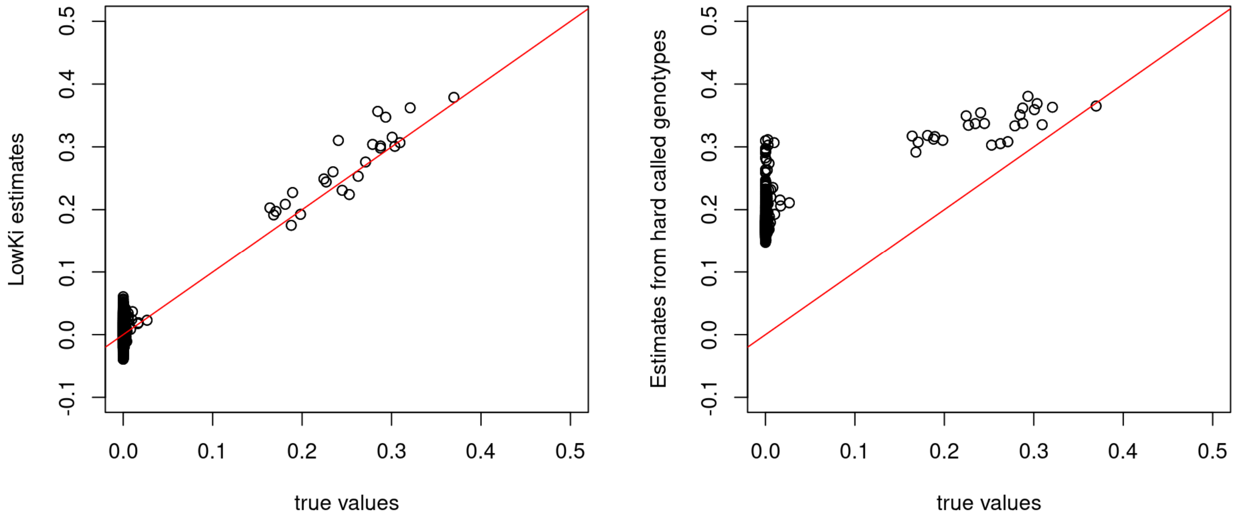


Fig 2. As Figure 1 but for estimation of the fraternity matrix, LowKi against GRM based on hard-called genotypes

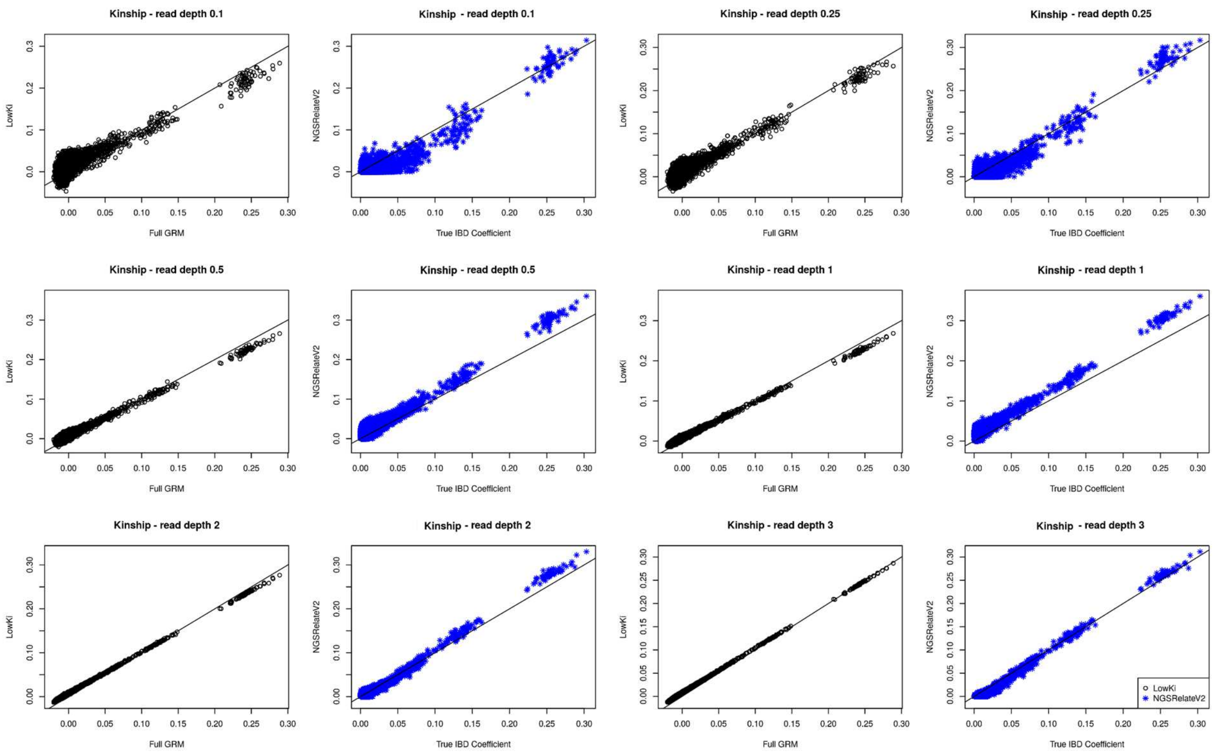


Fig 3. Estimation of the kinship matrix; LowKi against NGSRelateV2 for simulated data at low-depths. LowKi's estimates approach those that would have been attained from a 'Full GRM' - a GRM calculated using perfect simulated genotypes. NGSRelateV2 approaches the true IBD sharing coefficients that it aims to estimate.

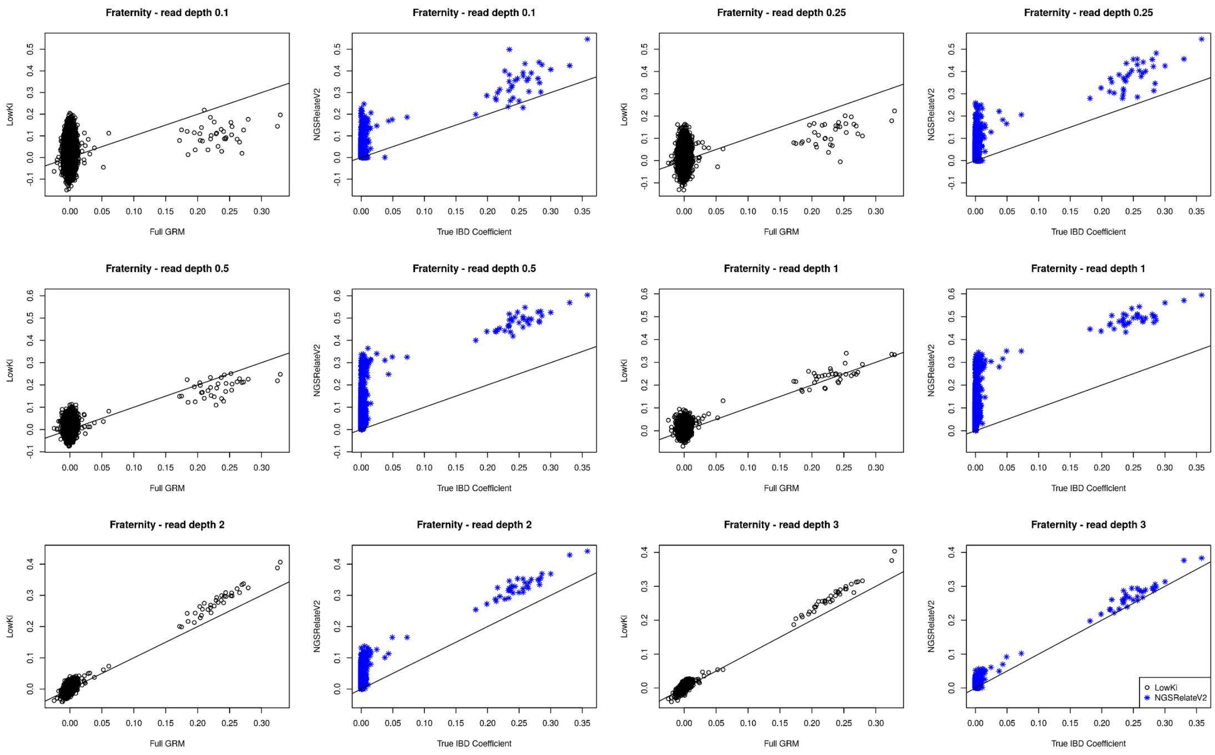


Fig 4. As Figure 3 but for the estimation of the fraternity matrix; LowKi against NGSRelateV2.

Acknowledgements

This work was supported by LABEX GENMED funded as part of “Investissement d’avenir” program managed by Agence Nationale pour la Recherche (grant number ANR-10-LABX-0013), and by the French regional council of Pays-de-le-Loire (VaCaRMe project). This work was also supported by the POPGEN project as part of the Plan Médecine Génomique 2025 (FMG2025/POPGEN) and by Inserm cross cutting project GOLD.

We would like to thank kindly all of the participants of the Vallo di Diano project and the Cilento cohort. We acknowledge the Center of Biological Resources (CHU Nantes, Hotel Dieu, CRB, Nantes, F-44093, France), the Dijon CRB, the CEPH and the Genomics and Bioinformatics Core Facility of Nantes (GenoBiRD, Biogenouest). We thank the FranceGenRef Consortium for giving us the opportunity to work with the FranceGenRef sequencing data for this project. We are also very grateful to Emmanuelle Génin for her valuable advice and insights regarding this project.

References

1. Novembre J, Johnson T, Bryc K, Kutalik Z, Boyko AR, Auton A, et al. Genes mirror geography within Europe. Nature. 2008;456: 98–101. doi:10.1038/nature07331
2. Jombart T, Devillard S, Balloux F. Discriminant analysis of principal components: a new method for the analysis of genetically structured populations. BMC Genetics. 2010;11: 94. doi:10.1186/1471-2156-11-94
3. Price AL, Zaitlen NA, Reich D, Patterson N. New approaches to population stratification in genome-wide association studies. Nat Rev Genet. 2010;11: 459–463. doi:10.1038/nrg2813
4. Yang J, Benyamin B, McEvoy BP, Gordon S, Henders AK, Nyholt DR, et al. Common SNPs explain a large proportion of the heritability for human height. Nature Genetics. 2010;42: 565–569. doi:10.1038/ng.608
5. Yang J, Lee SH, Goddard ME, Visscher PM. GCTA: a tool for genome-wide complex trait analysis. Am J Hum Genet. 2011;88: 76–82. doi:10.1016/j.ajhg.2010.11.011
6. Zhu Z, Bakshi A, Vinkhuyzen AAE, Hemani G, Lee SH, Nolte IM, et al. Dominance genetic variation contributes little to the missing heritability for human complex traits. Am J Hum Genet. 2015;96: 377–385. doi:10.1016/j.ajhg.2015.01.001
7. Herzig AF, Nutile T, Ruggiero D, Ciullo M, Perdry H, Leutenegger A-L. Detecting the dominance component of heritability in isolated and outbred human populations. Scientific Reports. 2018;8. doi:10.1038/s41598-018-36050-7
8. McPeek MS, Sun L. Statistical tests for detection of misspecified relationships by use of genome-screen data. Am J Hum Genet. 2000;66: 1076–1094. doi:10.1086/302800
9. Marenne G, Ludwig TE, Bocher O, Herzig AF, Aloui C, Tournier-Lasserve E, et al. RAVAQ : an integrative pipeline from quality control to region-based rare variant association analysis. Genetic Epidemiology. 2022; [In Press]. doi:10.1002/gepi.22450
10. McCombie WR, McPherson JD. Future Promises and Concerns of Ubiquitous Next-Generation Sequencing. Cold Spring Harb Perspect Med. 2019;9: a025783. doi:10.1101/cshperspect.a025783
11. Schwarze K, Buchanan J, Fermont JM, Dreau H, Tilley MW, Taylor JM, et al. The complete costs of genome sequencing: a microcosting study in cancer and rare diseases from a single center in the United Kingdom. Genetics in Medicine. 2020;22: 85–94. doi:10.1038/s41436-019-0618-7
12. Herzig AF, Ciullo M, Leutenegger A-L, Perdry H, FranceGenRef Consortium. Moment Estimators of Relatedness From Low-Depth Whole-Genome Sequencing Data. BMC Bioinformatics [In Review]. 2022. doi:10.21203/rs.3.rs-1109592/v1
13. DePristo MA, Banks E, Poplin R, Garimella KV, Maguire JR, Hartl C, et al. A framework for variation discovery and genotyping using next-generation DNA sequencing data. Nat Genet. 2011;43: 491–498. doi:10.1038/ng.806
14. Vitezica ZG, Legarra A, Toro MA, Varona L. Orthogonal Estimates of Variances for Additive, Dominance, and Epistatic Effects in Populations. Genetics. 2017;206: 1297–1307. doi:10.1534/genetics.116.199406
15. The 1000 Genomes Project Consortium. A global reference for human genetic variation. Nature. 2015;526: 68–74. doi:10.1038/nature15393
16. Hanghøj K, Moltke I, Andersen PA, Manica A, Korneliussen TS. Fast and accurate relatedness estimation from high-throughput sequencing data in the presence of inbreeding. Gigascience. 2019;8. doi:10.1093/gigascience/giz034
17. Dou J, Sun B, Sim X, Hughes JD, Reilly DF, Tai ES, et al. Estimation of kinship coefficient in structured and admixed populations using sparse sequencing data. PLOS Genetics. 2017;13: e1007021. doi:10.1371/journal.pgen.1007021
18. Browning BL, Browning SR. Genotype Imputation with Millions of Reference Samples. Am J Hum Genet. 2016;98: 116–126. doi:10.1016/j.ajhg.2015.11.020

Session 4, Amphi B

Functional and integrative genomics II

Single-cell RNA sequencing reveals aberrant megakaryocyte-erythroid progenitor and megakaryocyte populations in thrombocytopenic ETV6 variant carriers

Timothée Bigot[1], Laurent Hannouche[1], Elisa Gabinaud[1], Véronique Sbarra[1], Dominique Payet-Bornet[2], Marie Loosveld[2], Delphine Potier[2], Marie-Christine Alessi[1] and Marjorie Poggi[1]

[1] C2VN, 27 Bd Jean Moulin, 13005, Marseille, France
[2] CIML, 163 avenue de Luminy, 13288, Marseille, France

Corresponding Author: laurent.hannouche@univ-amu.fr

Abstract Background: *Germline mutations in the ETV6 gene are responsible for a familial thrombocytopenia and leukemia predisposition. Although previous studies showed that ETV6 plays an important role in megakaryocyte (Mk) maturation and platelet formation, mechanisms by which ETV6 dysfunction promotes thrombocytopenia remain unclear.* Goal: *To study the effect of ETV6 variations on the differentiation of hematopoietic-stem/progenitor-cells (HSPC) at a single cell level and discover changes mediating the differentiation defect.* Methods: *Single-cell RNA sequencing (scRNAseq) was performed on intermediate (day 6) and late (day 11) CD34^{+}-derived MKs issued from two ETV6 variant carriers (P214L and F417LfsTer4 variants). To analyze distinct subpopulations in the control and ETV6 mutant cells, dimension reduction was performed and cell clustering was visualized using UMAP reduction. Then clusters have been assigned to cell type and lineage signature gene sets were proposed.* Results: *The transcriptomic profile of ETV6 variants are similar and strongly differ from the control subjects, indicating that the same dysregulation is observed with both variations. We distinguished six distinct subpopulations in our model: HSPC, common-myeloid progenitors (CMP), MK-primed CMP, granulocyte-macrophage progenitors (GMP), megakaryocyte-erythroid progenitors (MEP), progenitor and mature MK (MkP/Mk), for which we propose molecular signature. ETV6 deficient condition is characterized by a higher proportion of HSPC and a reduced proportion of MkP/Mk. Genes differentially expressed between genotypes were identified for each hematopoietic cell stage. The transcriptomic profile of HSPC and CMP overlapped between control and patient cells. Pathway analysis using gene ontology identified the onset of divergence at the CMP stage, particularly for genes involved in the "regulation of protein localization in the nucleus". Divergence intensifies at MEP and MkP/Mk stages. Mitochondrial metabolism is the most significantly downregulated pathway in patients.* Conclusion: *ETV6 mutations affect early hematopoiesis, at the level of the CMP/MEP transition, resulting in an aberrant MEP population and a defect in megakaryopoiesis. Studying mitochondrial cellular metabolism may represent an interesting target to investigate. scRNAseq is a powerful technique through which we propose a systematic approach to unravel the mechanism of other inherited thrombocytopenias.*

Keywords ETV6, Thrombocytopenia, Megakaryopoieisis, scRNAseq, Oxydative-phosphorylation

1. Introduction

ETS-variant 6 (ETV6) is a repressor transcription factor (TF) of the ETS family, that holds 8 exons and encodes a 452 long amino acid protein. It is characterized by three functional domains. ETV6 binds to DNA via a C-terminal DNA-binding domain, highly conserved among ETS-family TFs. The ETV6 N-terminal Pointed (PNT) domain, also termed helix-loop-helix or sterile alpha motif domain mediates homo or heterodimerization with other TFs. The central domain regulates the repressor activity through binding of various co-repressors which subsequently recruit histone deacetylases to mediate transcriptional repression [1].

ETV6 was discovered in a leukemia-associated chromosomal translocation [2] and controls hematopoietic stem and progenitor cells (HSPC) function and thrombopoiesis [3,4]. Disruption of *Etv6* after the HSC commitment to the megakaryocyte/erythroid lineage resulted in consistent, marked reduction of platelet counts [4].

We previously characterized 6 heterozygous germline variations in the ETV6 gene leading to thrombocytopenia (THC5, OMIM 616216) with a predisposition to hematologic malignancies [5]. ETV6-related thrombocytopenia (ETV6-RT) is a highly penetrant form of inherited thrombocytopenia with autosomal dominant mode of inheritance. Since the first description of germline ETV6 mutation, 23 families (96 cases) and 16 ETV6 variants have been reported [6]. The common phenotype is a moderate thrombocytopenia sometimes associated with bleeding and with a predisposition to acute T or B lymphoblastic leukemia, B-cell lymphoma, acute myeloid leukemia or myelodysplasia [7]. All characterized ETV6 variants, reduced the ETV6 repressive activity due to modifications of ETV6 cellular localization, reduced DNA binding or epigenetic changes [8,9,10].

During megakaryopoiesis multipotent stem cells lead to a series of committed progenitors, the megakaryocytes (MK) being canonically derived through the HSPC, common myeloid progenitors (CMP), and megakaryocyte-erythroid progenitors (MEP). Alternative views have emerged in which a priming towards a particular lineage exists with MK-biased HSPC [11]. Presuming that ETV6 mutations have selective effects at a particular cell stage, we have used single cell RNA-sequencing (scRNAseq) to examine transcriptome of HSPC-derived MK issued from controls and ETV6 variant carriers: the newly identified ETV6 variant pF417LTer4 (NM_001987, c.1251del) and the previously described variant p.P214L (NM_001987, c.C641T) (figure 1). Using this method, we sought to understand underappreciated changes in gene expression in subpopulations of HSPCs and identify the molecular mechanisms of alterations in MK maturation resulting from germline ETV6 mutations.

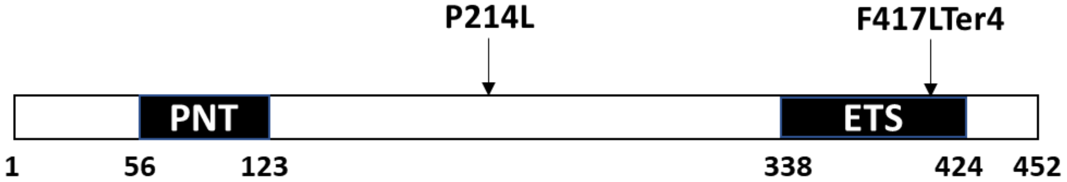


Figure 1. Schematic diagram of ETV6 protein with functional domains and the distribution of both germline mutations. PNT pointed domain (homo-oligomerization). ETS Ets DNA binding domains.

2. Methods

2.1. Single-cell RNA sequencing experiment

circulating isolated CD34+ cells from 2 healthy controls and 2 ETV6-variants carriers (P214L and F417LTer4) were cultured in StemSpan Serum-Free Expansion Medium II supplemented with Megakaryocyte Expansion Supplement (Stemcell Technologies) to induce MK differentiation. CD34+-derived differentiated cells were harvested at day 6 and 11 of the culture. Cell concentration and viability was assessed in automated cell counter (EveTM NanoEntek) and trypan blue staining. Cells were incubated with unique hashtag oligo-associated antibodies for each individual (HTO, TotalSeq, Biolegend), pooled and then finally single-cell isolation was achieved with the 10x Genomics System using the Chromium Next GEM Single Cell 5' Kit v2 (10X Genomics, ref 1000263), according to the manufacturer's protocol. Libraries were sequenced using a NextSeq 500 Illumina Technology (GBimmM platform).

2.2. Single-cell RNA sequencing data analysis

For both experiments, data preprocessing was performed by aligning mRNA library reads to the hg38 version of the human genome and quantified using 10x Genomics CellRanger count (version 6). Antibody counts for cell hashing were quantified using CITE-seq-count (version 1.4.3) with default parameters. Both mRNA & hashtags oligos (HTO) matrices were imported into R (version 4.1.2) and all downstream analyses were performed with the Seurat package (version 4.0.5) as previously described [12,13]. For both experiments, cells were demultiplexed using Seurat::HTODemux and Seurat::MULTIseqDemux functions. To confirm and enhance demultiplexing, Souporcell [14] (version 2.0.0) was performed because of its ability to identify specific single nucleotide variants (SNVs) for every patient sample. With both techniques, cell multiplets and negatives were removed from the filtered CellRanger barcode matrix. Before mRNA analyses, low quality cells (i.e. cells with less than 1000 mRNA and cells with more than a percentage of mitochondrial associated genes expression calculated by the function scater::isOutlier) and genes with poor information (i.e. genes expressed in less than 3 cells) were filtered out. Next, expression raw counts were normalized using Seurat::NormalizeData function ("LogNormalize" algorithm) and scaled with Seurat::ScaleData function. Samples from the two experiments were merged using the Seurat merge function. Integration by fixing anchors was not necessary, the quality of the merging and absence of batch effects were subsequently assessed by overlaps of several cell groups at both time points. After normalization and regression scaling (based on cell cycle associated genes), two dimensionality reductions were performed, the merged object was first reduced using principal component analysis (PCA) (Seurat::RunPCA) with the 1000 most variable genes (Seurat::FindVariableFeatures). The PCA reduction was used to compute a UMAP with the first 20 components (Seurat::RunUMAP). The nearest-neighbor graph was computed using Seurat::FindNeighbors function with a k parameter equal to 30 and the first 10 components. Cell clusters were identified with Seurat::FindClusters function with Louvain algorithm and using a resolution of 0.8. Based on the expression of celltype marker genes, non-myeloid cells were excluded (B and T cells, Mono/Macro) of the analysis. The Seurat analysis was performed again on this new subseted dataset using 1000 variable genes, default parameters scaling, nearest-neighbors graph with 10 components and k parameter equal to 30, clustering resolution of 0.8 and a UMAP construction with 28 components. For further analyses, this Seurat object was split between controls and ETV6 patients. Each subseted object was individually processed as before mentioned.

2.3. Differential gene expression (DEGs) and enrichment analysis

We performed DEGs analysis using Seurat::FindMarkers function for each pair of cell type controls versus ETV6 patients with these parameters: logfc.threshold equal to 0 and bimod test. DEGs upregulated and downregulated were separated following average log2 fold change and filtered by the p_value adjusted at 5%. Enrichment analysis were performed on this set of gene on 5 databases (Transfac and Jaspar, KEGG 2019, GO Biological process 2018, Reactome 2016 and Transcription factor PPI). Results were synthetized combining common genes and therefore removing redundant pathways.

3. Results

3.1. Cell type signatures

Firstly, cell type assignation was done based on supervised (public databases and literature) and unsupervised approach (top 10 DEGs by cluster). To avoid any biases due to ETV6 variations, this combined technique made possible the creation of 6 signatures composed of sets of genes in control object (table 1). Secondly, the suggested signatures were assigned to ETV6 patient object.

HSPC	CMP	CMP_Primed	GMP	MEP	MKP/MK
AVP	*MS4A3*	*LMNA*	*CSTA*	*CD38*	*MPIG6B*
PROM1	*RNASE2*	*CAVIN2*	*MPO*	*KLF1*	*PF4*
CHRBP	*CLC*	*LAT*	*ELANE*	*DEPTOR*	*GP9*
FLT3	*CSF2RB*	*VWA5A*	*PRTN3*	*ACSM3*	*VWF*
HOPX	*EPX*	*KRT1*	*AZU1*	*APOC1*	*PPBP*
IGHM	*CTSG*	*TMEM176B*	*RNASE3*	*TFRC*	*SH3BP5*
GLIPR		*KIT*	*CFD*	*ANK1*	*SELP*
		HPGDS		*TMSB10*	*LTBP1*
		TPSB2		*ALDH1A1*	*TMEM40*
		TPSAB1		*MT-ATP8*	*ITGA2B*
		SAMSN1		*MACROH2A1*	
		CD44		*GYPB*	

Table 1. Table of suggested signatures for cell type assignation

3.2. Cell types visualization and distribution

The previously described cell types were assigned to the merged object using the cell identity barcode and same cell types names. Cell types corresponding to immature cells (HSPC, CMP, GMP) were located on the left side of the figure (figure 2). For these early differentiating cells, transcriptomic profile of controls and ETV6 variants overlapped indicating similar profiles. On the right side were located the MEP and MkP/Mk cells. Interestingly, for these late stages of differentiation, cells from ETV6 patients exhibited distinct patterns of gene expression compared with controls. For each genotype, we defined the percentage of cell in each category (figure 3). The distribution of cells differs according the genotype. ETV6 deficient condition is characterized by a higher proportion of HSPC and a reduced proportion of MkP/Mk. To deepen the analysis, we listed the number of cells in each cell type for both conditions at day 6 and 11 to evaluate temporal differences. Notably, at day 6, on 2 391 and 4 044 cells processed respectively for control and ETV6 patients, the absolute numbers of HSPC were 126 vs 796 (4-fold increase) and of MkP/MK 97 vs 254 respectively in control and ETV6 patients. At day 11, on 2 071 et 1806 total cells, there were 12 control HSPC and 316 patient HSPC (60-fold increase) while there were 1617 control MkP/Mk and 663 patient MkP/Mk (2-fold decrease). Overall, these results indicate that ETV6 mutations affect early hematopoiesis, at the level of the CMP/MEP transition, resulting in an aberrant MEP population and a defect in megakaryopoiesis.

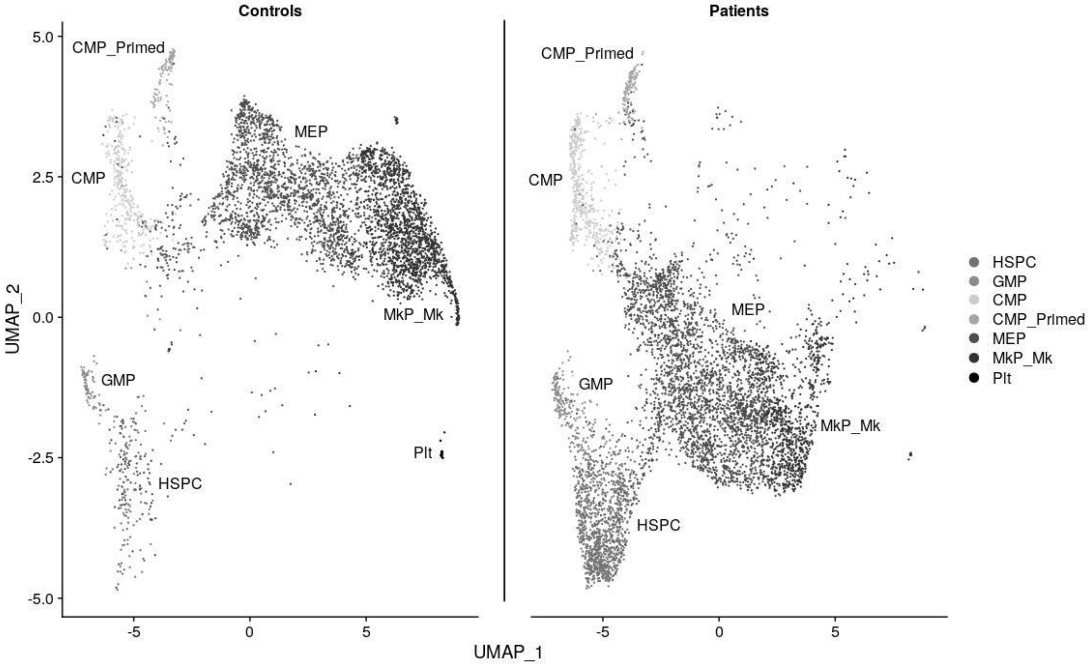


Figure 2. Cell type visualization on controls and ETV6 patients using UMAP reduction

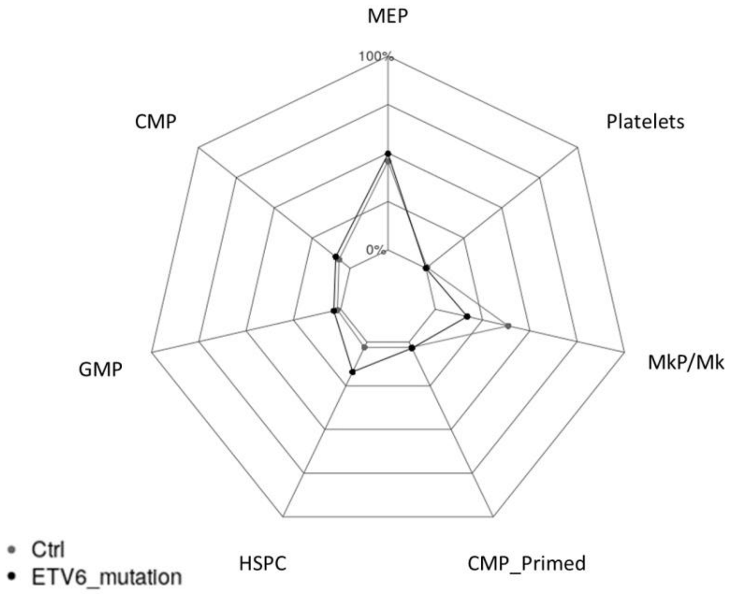
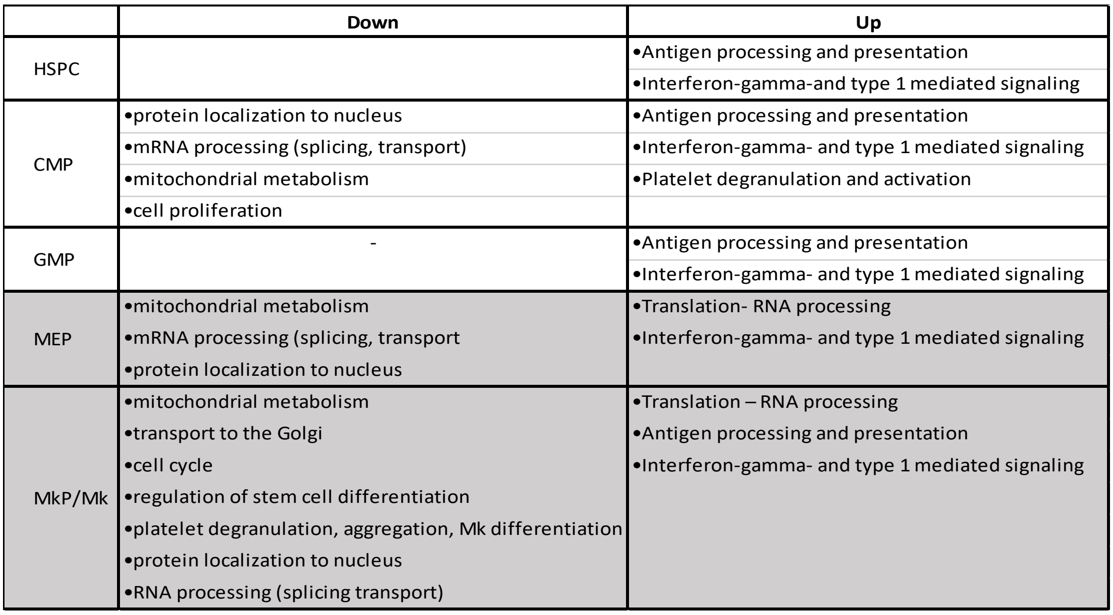


Figure 3. Cell proportion of each cell type in controls and ETV6 patients (days 6 and 11)

3.3. Enrichment analysis

The number of DEGs between control and ETV6 patient increased alongside the MK differentiation. Indeed, we identified 49 upregulated and 457 downregulated genes in ETV6 variant CMP, 225 upregulated and 424 downregulated genes in ETV6 variant MEP. and 463 upregulated and 1124 downregulated genes in ETV6 variant MkP/Mk. W Immature cells (i.e HSPC, CMP and GMP) did not exhibited a high number of DEGs. Enrichment analysis confirmed our previous observations and unrevealed various affected pathways in aforementioned cell types (table 2).



	Down	Up
HSPC		•Antigen processing and presentation •Interferon-gamma-and type 1 mediated signaling
CMP	•protein localization to nucleus •mRNA processing (splicing, transport) •mitochondrial metabolism •cell proliferation	•Antigen processing and presentation •Interferon-gamma- and type 1 mediated signaling •Platelet degranulation and activation
GMP	-	•Antigen processing and presentation •Interferon-gamma- and type 1 mediated signaling
MEP	•mitochondrial metabolism •mRNA processing (splicing, transport •protein localization to nucleus	•Translation- RNA processing •Interferon-gamma- and type 1 mediated signaling
MkP/Mk	•mitochondrial metabolism •transport to the Golgi •cell cycle •regulation of stem cell differentiation •platelet degranulation, aggregation, Mk differentiation •protein localization to nucleus •RNA processing (splicing transport)	•Translation – RNA processing •Antigen processing and presentation •Interferon-gamma- and type 1 mediated signaling

Table 2. Summary of affected pathways in ETV6 patients

4. Discussion

ScRNAseq experiment highlighted cell heterogeneity in our culture model catching numerous different transcriptomic profiles between cell types and different genotypes. Indeed, we identified 6 different cell types in both conditions (controls and ETV6 patients) thanks to our designed signatures (figure 2) whereas we could have expected to a homogeneous population after 11 days of culture. Among these 6 identified cell types, 5 are usual and typical of hematopoiesis like HSPC, CMP, GMP, MEP, MkP & Mk and one is rare, and not clearly described in the literature: Mk-primed CMP. In our model, alongside with the Mk-primed HSPC described population [11], this cell type could have the potential to directly differentiate from CMP to MkP or Mk and could represent a cell niche to rapidly produce Mk. Very interestingly, in controls only, some Mk were able to produce platelets indicating the presence of two differentiation profiles between our samples (not shown).

Profile differences between healthy control and ETV6 patients were quantified by analyzing the proportion of each cell type at days 6 and 11. The number of HSPC decreased in control condition at day 11 compared day 6 (10-fold decrease) but remained high in ETV6 patients, meaning that a defect is present in differentiation process associated with a higher proliferation rate in patient HSPC. This defect leads to a delayed and diminished differentiation for downstream cell types ending with MkP/Mk which is the most affected population. Indeed, the production of Mk was highly disrupted in patient condition relative to controls and suggest that their Mk are aberrant. These results are in accordance with published experimental data [5].

Quantitative analysis enabled detection of key populations in megakaryopoiesis defects but was not able to decipher the origin of deregulation and the mechanism affected during the process.

As previously described, megakaryopoiesis mainly occurs through a canonical differentiation going from HSPC, CMP, MEP and finally to MkP/Mk. In our culture model, we observed a main trajectory, which bypass the CMP population, leading to a direct differentiation from HSPC to MEP. This analysis was conducted by trajectory inference (data not shown) and was relatable with published data [11].

Regarding cell type distribution in UMAP (figure 3), we noticed that controls and patients shared close profiles for early populations like HSPC, CMP and GMP. For late populations like MEP and MkP/Mk, their distributions were totally separated meaning their transcriptomic profiles are different. In light of these observed transcriptomic differences, we could emphases genes involved in megakaryopoiesis defect in each cell type by focusing on differentially expressed genes. These DEGs increased from the CMP/MEP transition to the late differentiation stages, meaning an early role of ETV6 during the megakaryopoiesis. Enrichment analysis confirmed this result because the number of deregulated pathways increased across differentiation. These deregulations seem to start at CMP stage leading to abnormal CMP-MEP transition and resulting in MEP and MkP/Mk aberrant populations. One of the most significant deregulated pathways in MkP/Mk was mitochondrial metabolism and more specifically the Oxydative Phosphorylation (OP). Defect in OP was confirmed experimentally and studying mitochondrial cellular metabolism may represent an interesting target to investigate. Its phenotypic restoration could be an important step forward in restoration of Mk activity and role.

5. References

1. R. G. Lopez, C. Carron, C. Oury, P. Gardellin, O. Bernard, et J. Ghysdael, « TEL Is a Sequence-specific Transcriptional Repressor », *Journal of Biological Chemistry*, vol. 274, nº 42, p. 30132-30138, oct. 1999, doi: 10.1074/jbc.274.42.30132.
2. E. de Braekeleer *et al.*, « Identification of *NIPBL* , a new *ETV6* partner gene in t(5;12) (p13;p13)-associated acute megakaryoblastic leukemia », *Leukemia & Lymphoma*, vol. 54, nº 2, p. 423-424, févr. 2013, doi: 10.3109/10428194.2012.706288.
3. L. C. Wang *et al.*, « The TEL/ETV6 gene is required specifically for hematopoiesis in the bone marrow », *Genes Dev*, vol. 12, nº 15, p. 2392-2402, août 1998, doi: 10.1101/gad.12.15.2392.
4. H. Hock *et al.*, « Tel/Etv6 is an essential and selective regulator of adult hematopoietic stem cell survival », *Genes Dev.*, vol. 18, nº 19, p. 2336-2341, oct. 2004, doi: 10.1101/gad.1239604.
5. M. Poggi *et al.*, « Germline variants in ETV6 underlie reduced platelet formation, platelet dysfunction and increased levels of circulating CD34+ progenitors », *Haematologica*, vol. 102, nº 2, p. 282-294, 2017, doi: 10.3324/haematol.2016.147694.
6. J. Di Paola et C. C. Porter, « ETV6-related thrombocytopenia and leukemia predisposition », *Blood*, vol. 134, nº 8, p. 663-667, août 2019, doi: 10.1182/blood.2019852418.
7. S. Feurstein et L. A. Godley, « Germline ETV6 mutations and predisposition to hematological malignancies », *Int. J. Hematol.*, vol. 106, nº 2, p. 189-195, août 2017, doi: 10.1007/s12185-017-2259-4.
8. S. Topka *et al.*, « Germline ETV6 Mutations Confer Susceptibility to Acute Lymphoblastic Leukemia and Thrombocytopenia », *PLoS Genet*, vol. 11, nº 6, p. e1005262, juin 2015, doi: 10.1371/journal.pgen.1005262.
9. M. Y. Zhang *et al.*, « Germline ETV6 mutations in familial thrombocytopenia and hematologic malignancy », *Nat Genet*, vol. 47, nº 2, p. 180-185, févr. 2015, doi: 10.1038/ng.3177.
10. M. H. Fisher *et al.*, « ETV6 germline mutations cause HDAC3/NCOR2 mislocalization and upregulation of interferon response genes », *JCI Insight*, vol. 5, nº 18, p. e140332, sept. 2020, doi: 10.1172/jci.insight.140332.

11. B. Psaila et A. J. Mead, « Single-cell approaches reveal novel cellular pathways for megakaryocyte and erythroid differentiation », *Blood*, vol. 133, n^{o} 13, p. 1427-1435, mars 2019, doi: 10.1182/blood-2018-11-835371.
12. A. Butler, P. Hoffman, P. Smibert, E. Papalexi, et R. Satija, « Integrating single-cell transcriptomic data across different conditions, technologies, and species », *Nat Biotechnol*, vol. 36, n^{o} 5, p. 411-420, mai 2018, doi: 10.1038/nbt.4096.
13. T. Stuart *et al.*, « Comprehensive Integration of Single-Cell Data », *Cell*, vol. 177, n^{o} 7, p. 1888-1902.e21, juin 2019, doi: 10.1016/j.cell.2019.05.031.
14. H. Heaton *et al.*, « Souporcell: robust clustering of single-cell RNA-seq data by genotype without reference genotypes », *Nat Methods*, vol. 17, n^{o} 6, p. 615-620, juin 2020, doi: 10.1038/s41592-020-0820-1.

Session 4, Amphi C

Workflows, reproducibility & open science II

Bioinformatics platforms and data management : the GenOuest experience

Konogan BOURHY, Olivier COLLIN

Univ Rennes, Inria, CNRS, IRISA F-35000 Rennes, France

Corresponding Author: olivier.collin@irisa.fr

Abstract *The abstract of the paper (optional for short contributions) must be typeset in italic, with Times New Roman 11-point font. The left and right margins must be set to 3cm. 350 words maximum.*

Keywords Bioinformatics facilities, Data management, Open Science, FAIR data

1. Introduction

Bioinformatics platforms are privileged witnesses of the evolution of Biology. With technical innovations in data production, Biology is transforming into a Data Science. Bioinformatics platforms, initially computing power providers, are now concerned with data management in a context of increasing data production.

Created in 2001, the GenOuest platform is no exception to this phenomenon and had to implement various solutions throughout its history to support and promote good data management practices that are now required in many scientific projects.

2. A trajectory

Since its establishment in 2001, GenOuest has been gradually more and more involved in biological data management. It started with reference data management with BioMAJ (2007), then metadata management with the EMME project (Experimental Metadata Management Environment, 2010), before focusing on data hosted on the platform itself with the BioDataManager project.

Over the course of these projects, the challenge of data management has emerged as a key element of a platform's activity and GenOuest has embarked further on this path with the e-Biogenouest (2014) followed by CeSGO (CPER 2015) for the establishment of an environment for scientific projects and data management.

This initial focus on scientific data and its management made it possible to initiate the maDMP4LS project (machine actionable Data Management Plan for Life Sciences) in partnership with IFB and Inist, whose ambition is to be able to exploit the information contained in data management plans.

3. Different data oriented services

3.1. CeSGO

Adopting a vision focused on the life cycle of scientific projects as well as the life cycle of scientific data, the CeSGO project integrates scientific services with a bioinformatics infrastructure. The services include scientific collaboration, project management, data exchange and sharing. All services are based on open source solutions (Wordpress, Owncloud, Kanboard, SEEK). CeSGO's services create a one-stop shop where scientists will find all the research data management (RDM) tools for their needs. These services encourage the adoption of FAIR data principles and improve the reproducibility of in silico experiments.

Lessons learned during this project :

- Always be careful of the adoption process. If there is an adoption cost, or if it interferes with the current practices, users will be reluctant to use the new services. It also means that you have to support the change with communication and training.

- Services suitable for office work will be overwhelmed by bioinformatics usage. For example, Owncloud is having severe problems with bioinformatics data, the huge amount of files creating bottlenecks.

3.2. My and maDMP4LS

This project has two main areas of development: on the one hand we have DMP OPIDoR, a tool for writing data management plans developed by the OPIDoR team of Inist, and on the other hand we have My, the account and project management tool of the GenOuest platform, but which is also adopted by roscoff's bioinformatics platform. While DMP OPIDoR has begun a transition to actionable machine DMPs, the My tool must be adapted to be able to support the import of information from a data management plan in order to feed the descriptions of the projects for better monitoring and better management of the spaces made available to them. We are witnessing a paradigm shift since the data management plan, originally a document mainly intended for funders, becomes a tool for configuring computing and storage environments, and allows bioinformatics platforms to fully integrate into data lifecycle management, feeding on metadata to better meet the needs of researchers, and in turn enriching them to strengthen the application of the FAIR principles. The My tool (https://github.com/genouest/genouestaccountmanager) is at the heart of this evolution, offering multiple functionalities both for users of the platforms but also for the managers of these platforms. The tool makes it easy to manage accounts and their expiration as well as assignments to groups and projects. In addition, it is designed to easily interface with other local tools to provide additional information regarding quotas and their use, simplified management of SSH keys, mailing list management, etc. More recently, developments carried out in partnership with the IFB has also made it possible to add the first elements for the billing of services (CPU cost, storage cost).

4. Perspectives and conclusion

The implementation of collaborative services on top of a computing infrastructure is a controversial move. It is not directly related to bioinformatics but in the context of better data management practices, it can bring benefits to the users.

The first developments within the framework of the maDMP4LS project allowed the import of the DMP as part of a project request, associated with a request for storage space. The next step is to further exploit the information contained in the DMP to enable a configuration of bioinformatics environments that is more tailored to the real needs of applicants. In the perspective of interoperability with recherche.data.gouv, new additions will be necessary to contribute to data management.

Acknowledgements

CeSGO has been funded by CPER. maDMP4LS is funded by ANR. GenOuest thanks the Région Bretagne for its continuous financial support.

Linda: an interactive framework for experimental design and management of metadata

Benjamin Dartigues[1], Alexis Groppi[1,2], Sylvain Prigent[3], Yves Gibon[3], Cécile Cabasson[3]
and Macha Nikolski[1,2]

[1] Univ . Bordeaux, Centre de Bioinformatique de Bordeaux (CBiB), 33076 Bordeaux.

[2] Univ . Bordeaux, CNRS, IBGC, UMR 5095, 33077 Bordeaux

[3] UMR Biologie du Fruit et Pathologie, Université de Bordeaux, INRAE, F-33140 Villenave d'Ornon

{benjamin.dartigues, macha.nikolski, alexis.groppi}@u-bordeaux.fr

Keywords Metadata, plant phenotyping experiment, MIAPPE, FAIR, ISA-tab

https://services.cbib.u-bordeaux.fr/linda/

Efficient capture of environmental data and their integration with data from experimental read-outs is a prerequisite for a better understanding of how a given phenotype arises in response to external stimuli. While transcriptomics, proteomics or metabolomics provide information of unprecedented depth about the interactions between the genotype and the environment [1], it is essential to be able to link omics data to genetic and/or environmental data of the experiment. Being able to efficiently access and process experimental metadata is essential for the identification of the mechanisms responsible for the phenotypes or traits of interest via top-down or bottom-up modeling approaches.

However, proper processing, storage and integration of data and metadata is often neglected due, among other reasons, to the relative lack of standardization [2]. A way to overcome the difficulty of this task is to provide biologists with a versatile tool that (1) relies on standardized and machine-readable metadata formats to describe all aspects of experimental conditions, (2) allows import / export to centralized repositories such as MetaboLights (http://www.ebi.ac.uk/metabolights) and (3) can be seamlessly connected with other resources through by relying on FAIR principles for research software.

Based on the work originally proposed in XEML-Lab [3], Linda (pLant experImeNtal metaDatA) aims to extend and facilitate the process of capturing and storing metadata and is now a web-based framework developed to assist in the experimental design as well as the collection and linking of metadata for plant experiments.

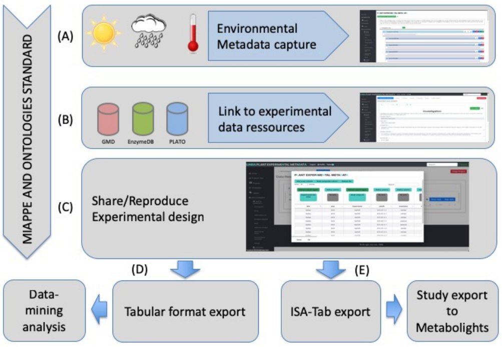

MIAPPE AND ONTOLOGIES STANDARD
(A)
Environmental Metadata capture
(B)
GMD
EnzymeDB
PLATO
Link to experimental data ressources
(C)
Share/Reproduce Experimental design
(D)
(E)
Data-mining analysis
Tabular format export
ISA-Tab export
Study export to Metabolights

Fig 1. Linda workflow. (A) Users can easily add new environmental metadata using dedicated ontologies or by uploading existing datasets in tabular format. (B) Link to remote databases is available to associate environmental metadata and experimental data. (C) Work can be saved and/or shared in order to complete or reproduce the experimental design. (D) Environmental and experimental data can be exported in tabular format for further analysis. (E) Export to ISA tab format is available.

Linda offers a complete tool for managing metadata associated with a phenotyping experiments on plant allowing, among other things, to describe experimental design and growth protocol in a FAIR manner using well-established ontologies (e.g., Plant Ontology, Crop Ontology, etc.) and dedicated norms such as the Minimal Information About Plant Phenotyping Experiment standard (MIAPPE) [4]. MIAPPE documents the minimal metadata necessary to ensure long term FAIRness of field or greenhouse datasets. The Investigation/Study/Assay (ISA) Metadata Framework for enabling discovery, exchange, and publication of metadata is also integrated and allows import/export in ISA-Tab format.

Linda was developed as a web application coupled with a graph-oriented database. This graph structure lends itself perfectly to the hierarchical structures described in the ISA tab and MIAPPE models. Linda allows the user to centralize their data and to have a complete overview of their experiments. Indeed, a user can describe his project and all the studies associated with this project while being guided and oriented by the MIAPPE standard. Within a study, he can accurately describe the biological materials used, with their origins (species, parent accession, etc.), their number, and their location in the experimental design. User is guided to describe its observed variables, its experimental factors as well as the biological materials used for your various studies. The use of a graph-oriented database in the background makes it possible to store your data in a secure and structured manner. The data collected will include environmental (climate, soil composition, nutrients, etc.), geographic (place of observation), temporal (time of observation), genotypic and phenotypic (from the whole plant to molecules) data. Linda can also capture any events that occurred during the experimentation time. The description of phenotypic, experimental and environmental variables is supported by observation units which group together all the observations made by the experimenter and which can be linked to spatial or temporal data. Any event occurring during the experience can also be notified and stored. Data from dataverse type repositories can be used upstream or downstream of the experiment. Downstream, experimental metadata (experimental design, growth protocol, biological material, etc.) and environmental (location details, latitude, altitude, field or greenhouse environment, etc.) can be used to build the complete structure of a project and underlying studies, describe the biological materials, even before possessing any measured data or metadata (observed variable as plant height, experimental data as glucose concentration).

Users can also document and enrich their project upstream of the experimentation by integrating, using the data loader, the data resulting from phenotypic and environmental observations. By integrating several models, Linda offers many formats to export, allowing interconnection between different suppliers (dataverse) and data integrators. Popular formats such as csv, tsv, json and excel are also available and allow easy data exchange and storage between users. By providing a user-friendly tool for both experimental design and management of metadata, and by enabling the linking of related experimental information to the metadata, Linda streamlines both the handling of metadata as well as its submission to central repositories such as MetaboLights. We believe that Linda offers an efficient tool for the integration of meta and experimental read-out data facilitating their exchange and upload to central repositories.

References

1. J K Nicholson, J C Lindon, and E Holmes. ’Metabonomics’: understanding the metabolic responses of living systems to patho- physiological stimuli via multivariate statistical analysis of biological NMR spectroscopic data. Xenobiotica; the fate of foreign compounds in biological systems, 29(11):1181–9, November 1999.

2. C Steinbeck et al. COordination of Standards in MetabOlomicS (COSMOS): facilitating integrated metabolomics data access. Metabolomics, 2015. ISSN 1573-3882.

3. J Hannemann et al. Xeml Lab: a tool that supports the design of experiments at a graphical interface and generates computer- readable metadata files, which capture information about genotypes, growth conditions, environmental perturbations and sampling strategy. Plant, cell & environment, 32(9):1185–200, September 2009. ISSN 1365-3040

4. C Pommier et al. *Data standards for plant phenotyping: MIAPPE and its implementations.* In *Proceedings Plant and Animal Genome XXVI Conference. San Diego* 2018

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

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, small 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 and 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 Copy Number Variation (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 python script and SnpEff

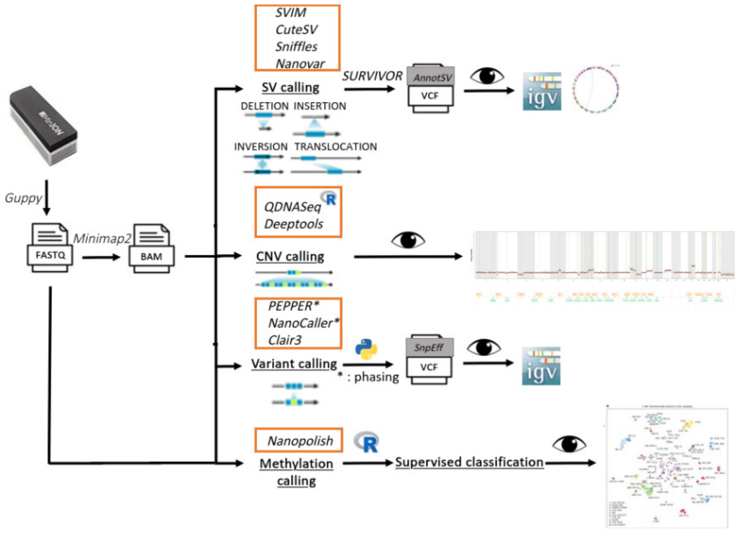


Fig. 1. Workflow of NanoCliD toolkit

annotates the resulting VCF. Variants are then reviewed manually with IGV. Finally, methylation calling based on FastQ and BAM files is achieved with Nanopolish. The output is used for supervised classification allowing t-SNE and UMAP visualization to assign the methylation profile to a class of brain tumors or sarcomas.

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. The identification of a tandem duplication thus contributed to improved 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 have been processed by 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 small 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] P. Euskirchen, F. Bielle, K. Labreche, et al. Same-day genomic and epigenomic diagnosis of brain tumors using real-time nanopore sequencing. *Acta Neuropathologica*, (134):691–703, 2017.

[2] Y. Sakamoto, S. Sereewattanawoot, and A. Suzuki. A new era of long-read sequencing for cancer genomics. *Journal of Human Genetics*, (65):3–10, 2020.

[3] Y. Wang, Y. Zhao, A. Bollas, 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] M. Kulis and M. Esteller. DNA methylation and cancer. *Advances in Genetics*, (70):27–56, 2010.

Session 5, Amphi A

Structural bioinformatics & proteomics II

KinDock 2.0: Un outil de criblage virtuel des protéines-kinases enrichi par l'intelligence artificielle.
KinDock 2.0: A robust protein-kinases virtual screening tool enhanced by machine learning predictors.

Victor Reys[1], Jean-Luc Pons[1], Violaine Moreau[1] and Gilles Labesse[1]
[1]Centre de Biologie Structurale, 29 Rue Navacelles, 34080, Montpellier, France

Corresponding author: labesse@cbs.cnrs.fr

Abstract *Despite many years of research in drug developments, cancer chemotherapy often fails due to various side-effects. Various approaches address this issue, including high-throughput profiling of the various compounds active against the protein-kinases. To complement them, we developed an integrated approach combining experimental data, structural bioinformatics and chemoinformatics tools with machine-learning to build dedicated models for affinity predictions on a set of 143 human protein-kinases.*

Keywords Protein-kinases, Structure-based virtual screening, Ligand-Based virtual screening, Machine-Learning, In Silico Profiling.

1 Introduction

Despite years of intensive research and development, cancer remains one of the leading causes of death worldwide. Chemotherapy is the most commonly used treatment for cancer, as surgery and radiation are often not effective against metastases. However, drug development often fails due to improper activity profile that induces side-effects and/or toxicity. So we need more precise guides to improve chemical compounds, especially for those targeting the large family of protein-kinases (∼520 members in humans). These important therapeutic targets are involved in numerous signaling cascades to control various biological processes (e.g.: cell proliferation) and many are oncogenic (e.g.: B-Raf) while some others act as tumor suppressors (e.g.: Csk). The most successful protein-kinase inhibitors (PKIs) are specific (e.g.:imatinib[1,2]) while some advert effects were due to lack of specificity (e.g.: unwanted targeting of the receptor VEGFR leading to heart failure). On the contrary, it would be interesting to target simultaneously a given set of protein-kinases along a signaling cascade (e.g.: B-Raf/MEK/Erk [3,4]) or in parallel pathways (e.g.: Akt and Erk) for enhanced efficiency and lower resistance. Today, this is achieved by bitherapies (with two different drugs) which are less optimal (due to possible drug-drug interactions). Unfortunately, all the protein-kinases share significant sequence and structure similarities, especially in their ATP-binding sites, which dramatically increase the difficulty to design new compounds that are more selective. So, we need new chemoinformatics and bioinformatics tools for accurate profiling drug candidates [5].

In this project, we wish to develop a new *in silico* tool dedicated to the profiling of ATP-competitive protein-kinases inhibitors. To this end, we developed a systematic integrated approach combining experimental data, structural bioinformatics and chemoinformatic tools with advanced machine-learning to build dedicated models for affinity predictions. This approach was prior validated on nuclear receptors [6]. Those models will help in both selecting appropriate compounds for chemotherapy as well as guiding medicinal chemists for the design of new therapeutic solutions. First, we gathered databases of experimental affinities, then built 3D structural models of each protein-kinase, performed systematic virtual screenings, and built machine-learning models for affinity predictions.

2 Material and Methods

2.1 Experimental Affinities Data-sets

Two independent and complementary databases of protein-kinases inhibitors have been gathered for machine learning and extensive testing.

Navigating the Kinome: 3 858 compounds (homogeneous dataset with different scaffolds) tested over 172 protein-kinases, with ∼36 000 experimental affinities (as Ki)[7].

Klaeger: 242 compounds (marketed or in clinical phase) over 219 protein-kinases, with 5 123 experimental affinities (as Kd)[8].

BindingDB: ∼ 50 000 compounds over 486 protein-kinases (heterogeneous dataset, from various laboratories), ∼ 180 000 experimental affinities (as Ki, Kd, IC50 and EC50)[9].

Prior to the docking procedure, compounds 3D conformations were modeled (from SMILES) using two standard software (Frog2[10] and OpenBabel[11]) and computed for each, two distinct atomic partial charges models (MMFF94 and Gasteiger) leading to 4 representations.

2.2 Structure-Based approaches

For 486 protein-kinases, 3D models were built using our server @TOME[12]. Generated models were well evaluated (score ≥0.4) using the Qmean[13] metric, which is a linear combination of terms from four different statistical potentials and often used in structure evaluation in the PDB[14]. This database is freely accessible (http://atome.cbs.cnrs.fr/kinome). For most protein-kinase, more than 20 useful templates can be detected. The diversity of templates allows us to model and sample multiple conformations of each target.

During this modeling step, ligands within the ATP-binding sites were extracted to serve as shape restraints (or anchors) to guide subsequent ligand comparative docking. Then, using the top 20 most similar anchors (from 2D ECFP4 Tanimoto similarity metric) to select the most compatible active sites, a set of 20 poses per ligand can be computed using PLANTS[15]. Hence, docking is prioritized into templates known to bind similar ligands and more likely to adopt compatible conformation. This is important in the case of the protein-kinases well known to adopt distinct conformations such as the well known DFG-in and DFG-out conformations.

Poses were evaluated using state-of-the-art scoring functions (DSXScore[16], Xscore[17], PLPChem[15] et MedusaScore[18]) as well as a consensus function (@TOMEScore) based on these four scoring functions.

A total of 19 structure-based descriptors can be retrieved from the modeling/docking study by @TOME which are presented in Table 1.

For each ligand, we computed the average and standard deviation of each descriptor over its 20 poses. This leads to a 38-long vector of structure-based descriptors per ligands.

2.3 Ligand-Based approaches

On the track of a previously described method by Schneider *et. al.,* [6], we included a set of 11 molecular property descriptors (described in detail in Table 2), generated from the RDKit library under R, that easily provide those descriptors computation from molecular structure in a multi-molfile. The selected Ligand-Based descriptors aims to complement the description of ligands by adding solvation, rigidity and volume information, that are not taken into account in the force-field potentials implemented in scoring functions.

2.4 Machine-Learning

Random Forest approaches were used to generate affinity prediction models. Based on a methods originally described and applied onto a nuclear receptor[6], it combines both ligand-based and structure-based information. The method was adapted for its systematic use on human protein-kinases. The learning was performed with 1000 trees, fitting the log(affinity) values by minimizing the mean average error (MAE). Performances were monitored under a 10-fold cross-validation process. The data-set was partitioned in 80% and 20% for the training and testing set respectively, taking care of the equal spread of affinities in each sets. All the models were built using the 'caret' package under R.

Name	Description
QMean	QMean score of receptor model[13]
PlantsRaw	PLANTS ChemPLP score[15]
Plants	PLANTS pKa (calculated by linear regression on PDBbind)
MedusaRaw	Medusa[18] original score
MedusaSc	MedusaScore pKa (calculated by linear regression on PDBbind)
XScore	XScore[17] affinity score (pKa)
DSXRaw	DSX [16] original score
DSXScore	DSX pKa (calculated by linear regression on PDBbind)
AtomeSc	@TOME [12] consensus pKa prediction (Avg. PLANTS, XScore, MedusaScore and DSX scores)
LigEff	Ligand efficiency (pKa / number of atoms)
Cosian	Avg. of ECFP2, ECFP4 and ISIDA tanimoto similarities
Tanimoto	Similarity between candidate ligand and anchor (ECFP4)[11]
NbAtLig	Number of atoms in candidate ligand
AnchorFit	Candidate/ligand superimposition score (PLANTS software)
LPC	LPC software score (receptor/ligand complementarity function)
Reliab	Ligand Position Error (SVM multi-variable linear regression)[12]
PSim	Similarity to receptor/ligand interaction profile in PDB template [12]
AtomSA	Secondary Assignment @TOME score [12]
CpxQuality	Complex quality consensus score [12]

Tab. 1. Set of 19 structure-based descriptors, describing docking pose

Name	Description
MW	Molecular weight
VABC	Volume descriptor
nAtom	Number of atoms
nAromBond	Number of aromatic bonds
nHBDon	Number of hydrogen-bounds donor
nHBAcc	Number of hydrogen-bounds acceptor
TopoPSA	Topological Polar Surface Area
XLogP	LogP prediction
HybRatio	Fraction of sp3 / sp2 carbons
nRotB.nB	Fraction of rotatable bonds / number of bonds
nAromB.nB	Fraction of aromatic bonds / number of bonds

Tab. 2. Set of 11 ligand-based descriptors used to describe molecules

3 Results

3.1 Docking campaign

The docking procedure yielded in ∼10% of correct poses (according to visual inspection) and a low overall correlation (R^2∼0.04) of the consensus scoring with experimental affinities. This demonstrates the limitation of the current methods and the need for machine learning to improved efficiency of the predictions. For the database Navigating the Kinome, this primary docking campaign led to ∼2.8 million poses, to be used in machine-learning.

3.2 Affinity predictions models

Regarding the experimental data available for each targets (Figure: 1), we see an unbalanced spread of the number of experimental affinities per target. This could lead to strong variations in robustness and performances for the random forest models during training. Accordingly, the training was first focused on a subset of targets with enough information available (templates and affinities).

Random forest affinity prediction models have been learned for 143 distinct protein-kinases and showed overall good performances with 0.7 RMSE (root mean squared error from predicted and measured affinities) and a good agreement with R^2∼0.33 (Figure: 2).

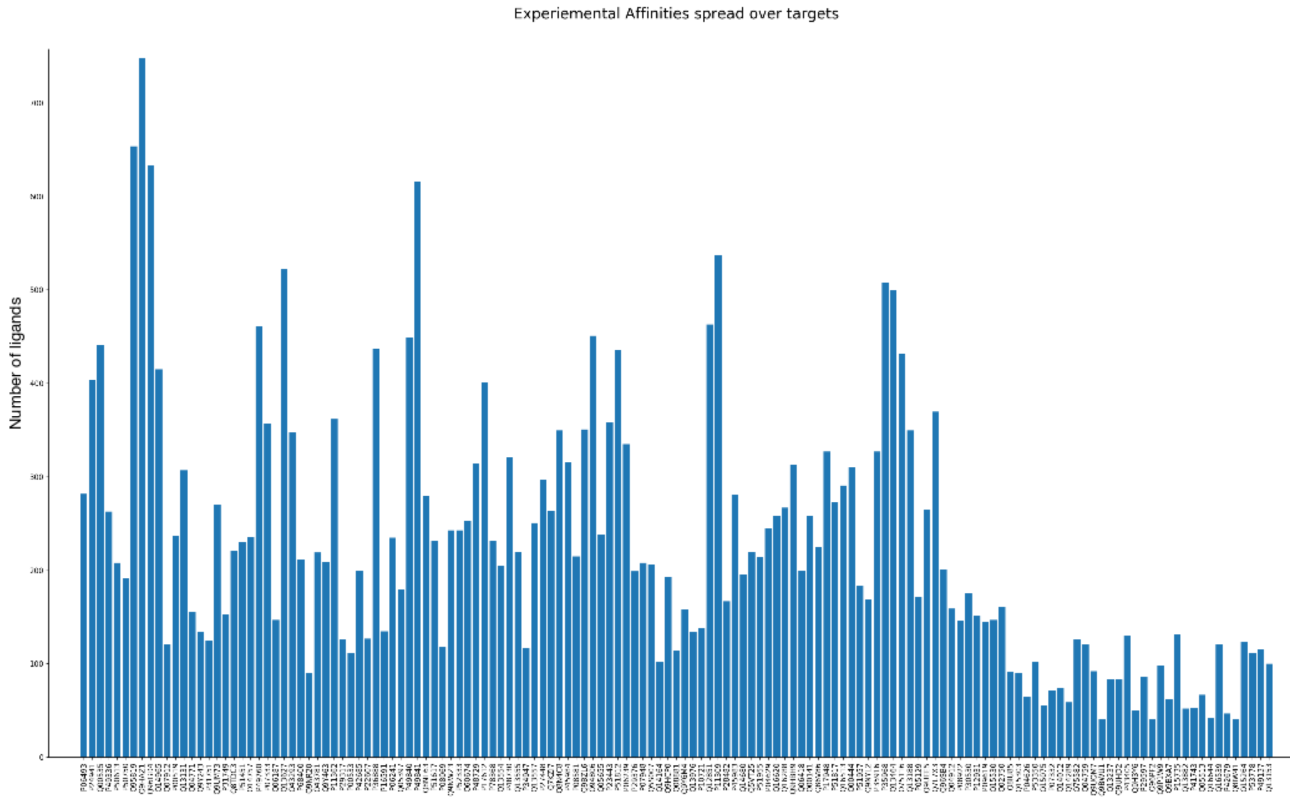


Fig. 1. Spread of the ligands over the 143 random forest models

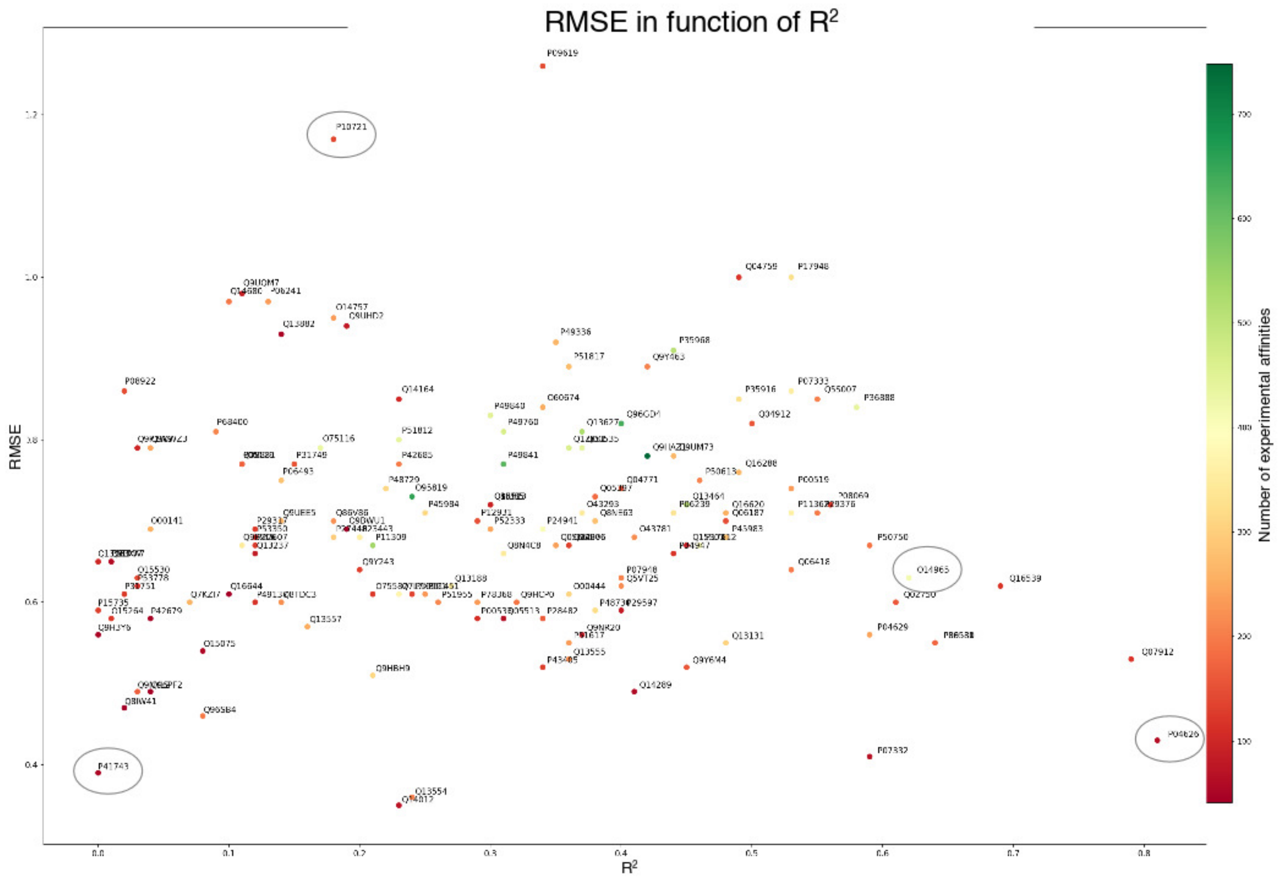


Fig. 2. Graphical representation of the two performances metrics (RMSE and R^2) for the 143 models. Targets analysed in Figure 3 are circled in grey.

However, some models showed sub-optimal performances. In Figure 3, we present 4 cases (that are circled in Figure 2), that represent the extreme cases of performance variability. The variability are mainly due to a low number of ligands, a specificity for the target and affinities data either limited to small ranges or unequally spread.

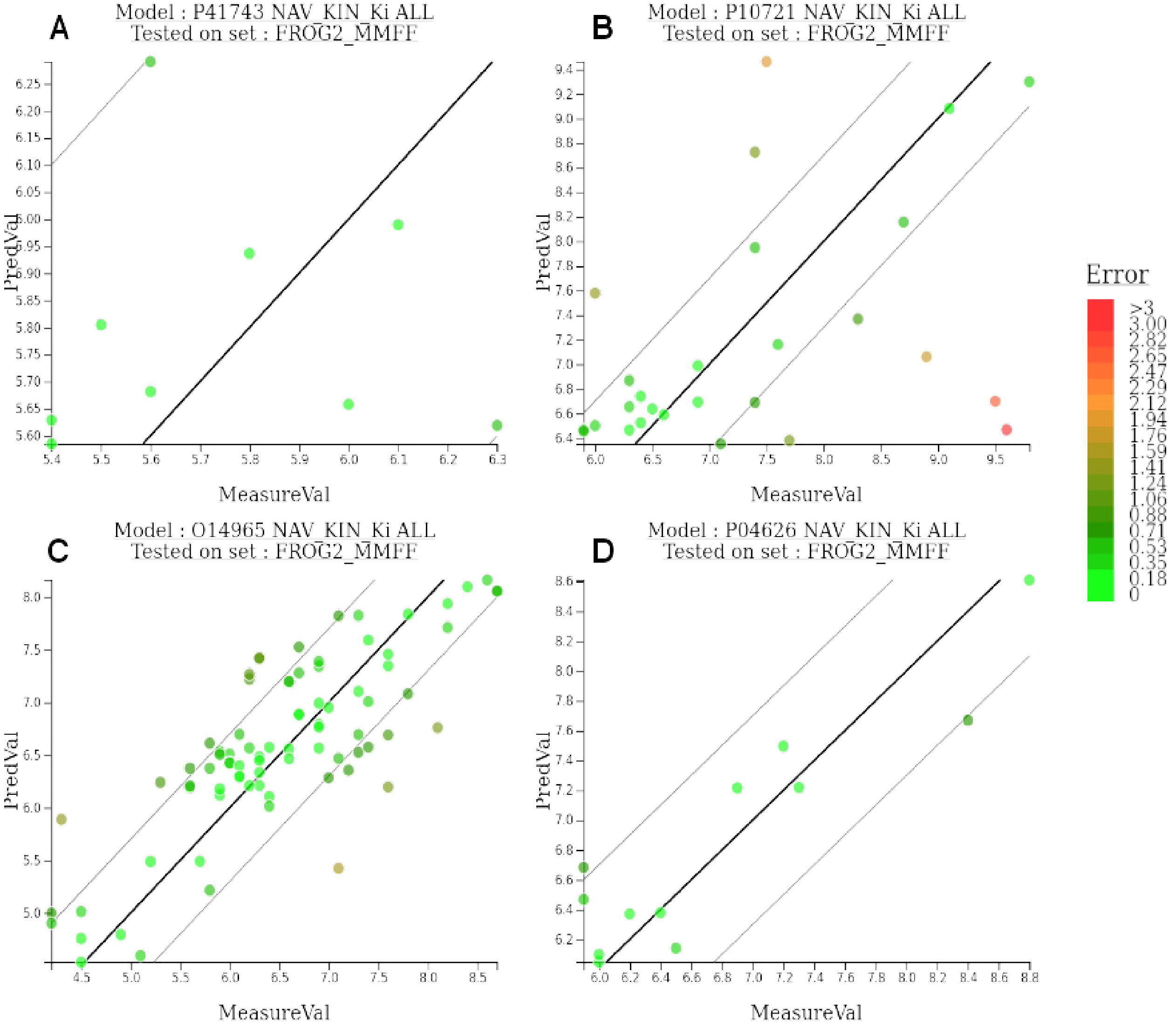


Fig. 3. A) KPCI predictions of 9 test-set ligands. The performances show a low RMSE of 0.38 and a R^2 of 0.0. B) KIT predictions of 27 test-set ligands. The performances show a high RMSE of 1.15 and is partially due to the absence of DFG-in templates. C) AURKA predictions of 81 test-set ligands. A typical target where most metrics went well, with a RMSE of 0.62 and a R^2 of 0.64. The model was trained with 334 ligands, which displayed no specific scaffold for high affinity ligands and a good spread of the affinities ranging from 4.2 to 9.4. D) ERBB2 predictions of 12 test-set ligands. The model RMSE of 0.4 and a R^2 of 0.85.

The serine/threonine kinase KPCI predictions of only 9 test-set ligands. The performances show a low RMSE of 0.38 which is better than the experimental variations. The low R^2 of 0.0 is linked to the low spread in the experimental affinities. In the case of the non-receptor tyrosine-kinase KIT predictions of 27 test-set ligands. The performances show a high RMSE of 1.15. This is due to the specificity of this target, that favorably binds a specific chemo-type scaffold, and the reason why high affinity ligands may be wrongly predicted, as the training set contains only few of these scaffold and the model was unable to understand this specificity. As an example of a very well behaved model, the predictions for the serine/threonine kinase AURKA on test-set of 81 ligands showed a rather low RMSE of 0.62 and a good R^2 of 0.64. The model was trained with 334 ligands with a good spread of the affinities ranging from 4.2 to 9.4 and variety in chemotypes. Finally for the receptor tyrosine-kinase ERBB2, our predictions of 12 test-set ligands. A typical example where performance metrics do no display prediction errors with a RMSE of 0.4 and a R^2 of 0.85. Affinity spread for this target ranges from 5.3 to 8.8, and the model was able to reproduce this spread on the test set. Nevertheless, the poor number of ligands within the training and testing set do not allow us to validate the robustness of this prediction model.

Although those models are generally well performing, we detected some limitations: a minimum of 150 experimental affinities per targets are required for guaranteed robustness and performances. Without enough information, affinity spreads, chemical diversity and training set size may be the limiting factors. Therefore, this limits this approach to well-studied targets.

4 Discussion

On the one hand, we built affinity predictions models are quite reliable, and therefore these models can be used for robust prediction of affinities for the 143 protein-kinases for which a model was trained. For the 143 random forest models, the average performances of 0.7 RMSE on test set is promising. Nevertheless, the average R^2 of ∼0.33 shows that RMSE should not be the only metric to optimize.

Up to now, predictions models were trained and tested using the Navigating the Kinome dataset[7]. For better validation, we are planning to employ supplementary datasets for external validation using data coming from the experimental screenings performed by Klaeger[8] and from BindingDB[9] which are currently being screened using our docking protocol.

5 Conclusion

We demonstrated that random forests can be successfully applied for affinity predictions on protein-kinases. The generated prediction models have on average an acceptable RMSE of ∼0.7, which fall in the range of experimental variability for a subset of well-characterized protein-kinases. We are planning to use other machine-learning tools, such as neural network approaches for the prediction of affinities, which could lead to better performances and could be applied on the complete human kinome.

We will provide, in the near future, a complete webserver tool for protein-kinases profiling, enhanced with dedicated machine-learning models for affinity prediction. This tool will hopefully help in selecting appropriate compounds for chemotherapy as well as guiding medicinal chemists for the design of new therapeutic solutions.

Acknowledgements

The authors thanks the ChemBioFrance structure for providing us the hardware able to process this project. We also have a special thanks and acknowledgement to la Ligue contre le Cancer for supporting Victor Reys with a PhD grant for this project.

References

[1] E. Buchdunger, J. Zimmermann, H. Mett, T. Meyer, M. Müller, B. J. Druker, and N. B. Lydon. Inhibition of the Abl protein-tyrosine kinase in vitro and in vivo by a 2-phenylaminopyrimidine derivative. *Cancer Res*, 56(1):100–104, Jan 1996.

[2] B. J. Druker, S. Tamura, E. Buchdunger, S. Ohno, G. M. Segal, S. Fanning, J. Zimmermann, and N. B. Lydon. Effects of a selective inhibitor of the Abl tyrosine kinase on the growth of Bcr-Abl positive cells. *Nat Med*, 2(5):561–566, May 1996.

[3] M. E. Lacouture, A. Desai, K. Soltani, V. Petronic-Rosic, A. E. Laumann, M. J. Ratain, and W. M. Stadler. Inflammation of actinic keratoses subsequent to therapy with sorafenib, a multitargeted tyrosine-kinase inhibitor. *Clin Exp Dermatol*, 31(6):783–785, Nov 2006.

[4] L. Adnane, P. A. Trail, I. Taylor, and S. M. Wilhelm. Sorafenib (BAY 43-9006, Nexavar), a dual-action inhibitor that targets RAF/MEK/ERK pathway in tumor cells and tyrosine kinases VEGFR/PDGFR in tumor vasculature. *Methods Enzymol*, 407:597–612, 2006.

[5] V. Reys and G. Labesse. In silico profiling of protein kinases inhibitors. *Med Sci (Paris)*, 36 Hors série n° 1:38–41, Oct 2020.

[6] M. Schneider, J. L. Pons, G. Labesse, and W. Bourguet. In Silico Predictions of Endocrine Disruptors Properties. *Endocrinology*, 160(11):2709–2716, 11 2019.

[7] J. T. Metz, E. F. Johnson, N. B. Soni, P. J. Merta, L. Kifle, and P. J. Hajduk. Navigating the kinome. *Nat Chem Biol*, 7(4):200–202, Apr 2011.

[8] S. Klaeger, S. Heinzlmeir, M. Wilhelm, H. Polzer, B. Vick, P. A. Koenig, M. Reinecke, B. Ruprecht, S. Petzoldt, C. Meng, J. Zecha, K. Reiter, H. Qiao, D. Helm, H. Koch, M. Schoof, G. Canevari, E. Casale, S. R. Depaolini, A. Feuchtinger, Z. Wu, T. Schmidt, L. Rueckert, W. Becker, J. Huenges, A. K. Garz, B. O. Gohlke, D. P. Zolg, G. Kayser, T. Vooder, R. Preissner, H. Hahne, N. Tõnisson, K. Kramer, K. Götze, F. Bassermann, J. Schlegl, H. C. Ehrlich, S. Aiche, A. Walch, P. A. Greif, S. Schneider, E. R. Felder, J. Ruland, G. Médard, I. Jeremias, K. Spiekermann, and B. Kuster. The target landscape of clinical kinase drugs. *Science*, 358(6367), 12 2017.

[9] M. K. Gilson, T. Liu, M. Baitaluk, G. Nicola, L. Hwang, and J. Chong. BindingDB in 2015: A public database for medicinal chemistry, computational chemistry and systems pharmacology. *Nucleic Acids Res*, 44(D1):D1045–1053, Jan 2016.

[10] M. A. Miteva, F. Guyon, and P. Tufféry. Frog2: Efficient 3D conformation ensemble generator for small compounds. *Nucleic Acids Res*, 38(Web Server issue):W622–627, Jul 2010.

[11] N. M. O'Boyle, M. Banck, C. A. James, C. Morley, T. Vandermeersch, and G. R. Hutchison. Open Babel: An open chemical toolbox. *J Cheminform*, 3:33, Oct 2011.

[12] J. L. Pons and G. Labesse. @TOME-2: a new pipeline for comparative modeling of protein-ligand complexes. *Nucleic Acids Res*, 37(Web Server issue):W485–491, Jul 2009.

[13] P. Benkert, M. Biasini, and T. Schwede. Toward the estimation of the absolute quality of individual protein structure models. *Bioinformatics*, 27(3):343–350, Feb 2011.

[14] S. K. Burley, C. Bhikadiya, C. Bi, S. Bittrich, L. Chen, G. V. Crichlow, C. H. Christie, K. Dalenberg, L. Di Costanzo, J. M. Duarte, S. Dutta, Z. Feng, S. Ganesan, D. S. Goodsell, S. Ghosh, R. K. Green, V. Guranović, D. Guzenko, B. P. Hudson, C. L. Lawson, Y. Liang, R. Lowe, H. Namkoong, E. Peisach, I. Persikova, C. Randle, A. Rose, Y. Rose, A. Sali, J. Segura, M. Sekharan, C. Shao, Y. P. Tao, M. Voigt, J. D. Westbrook, J. Y. Young, C. Zardecki, and M. Zhuravleva. RCSB Protein Data Bank: powerful new tools for exploring 3D structures of biological macromolecules for basic and applied research and education in fundamental biology, biomedicine, biotechnology, bioengineering and energy sciences. *Nucleic Acids Res*, 49(D1):D437–D451, 01 2021.

[15] O. Korb, T. Stützle, and T. E. Exner. Empirical scoring functions for advanced protein-ligand docking with PLANTS. *J Chem Inf Model*, 49(1):84–96, Jan 2009.

[16] G. Neudert and G. Klebe. DSX: a knowledge-based scoring function for the assessment of protein-ligand complexes. *J Chem Inf Model*, 51(10):2731–2745, Oct 2011.

[17] R. Wang, L. Lai, and S. Wang. Further development and validation of empirical scoring functions for structure-based binding affinity prediction. *J Comput Aided Mol Des*, 16(1):11–26, Jan 2002.

[18] S. Yin, L. Biedermannova, J. Vondrasek, and N. V. Dokholyan. MedusaScore: an accurate force field-based scoring function for virtual drug screening. *J Chem Inf Model*, 48(8):1656–1662, Aug 2008.

Template for JOBIM 2022 Highlight [Effect of quercetin on lipid membrane rigidity: assessment by atomic force microscopy and molecular dynamics simulations]

Jad EID[1,2], Alia JRAIJ[2], Hélène GREIGE-GERGES[2] and Luca MONTICELLI[2]

[1] Bioactive Molecules Research Laboratory, Doctoral School of Sciences and Technologies, Lebanese University, Beirut, Lebanon

[2] Molecular Microbiology and Structural Biochemistry (MMSB), CNRS & Univ. Claude Bernard Lyon I, UMR 5086, Lyon F-69007, France

Corresponding Author: Luca.monticelli@ibcp.fr

Paper Reference: **Eid et al. (2021) Effect of quercetin on lipid membrane rigidity: assessment by atomic force microscopy and molecular dynamics simulations, BBA Advances, 2021, https://doi.org/10.1016/j.bbadva.2021.100018**

Quercetin (3,3′,4′,5,7-pentahydroxyl-flavone) is a natural flavonoid with many valuable biological effects [1], but its solubility in water is low [2], posing major limitations in applications. The encapsulation of quercetin in liposomes could overcome the drawbacks related to the physico-chemical properties of this flavonoid [3]. However, when encapsulating drugs or other small molecules, it is important that liposome mechanical stability and elastic properties are preserved, since liposome elasticity impacts size, shape, membrane permeability, and drug loading efficiency [4,5]. Liposome elasticity may be affected by the incorporation of quercetin, but the effect of quercetin on liposome's elasticity is still not known. In this article, we quantified the effect of quercetin molecules on the rigidity of lipoid E80 liposomes using atomic force microscopy (AFM) and molecular dynamics (MD) simulations. In simulations, we used model membranes similar to the ones used in the AFM experiments, at the same temperature, to mimic AFM experimental conditions. AFM images show no effect of quercetin molecules on liposomes morphology and structure. However, AFM force curves suggest that quercetin softens lipid membranes; the Young modulus measured for liposomes encapsulating quercetin is smaller than that determined for blank liposomes. We then used MD simulations at atomistic level to interpret the effect of quercetin on membrane rigidity in terms of molecular interactions. The decrease in membrane rigidity was confirmed by the simulations, which also revealed that quercetin affects structural and dynamic properties: membrane thickness is decreased, acyl chains disorder is increased, and diffusion coefficients of lipid molecules are also increased. Such changes can be interpreted in the light of quercetin localization and orientation in the membrane: due to its amphipathic nature, quercetin is located preferentially at the interface between the hydrophobic core and polar head groups of the lipids, roughly parallel to the membrane normal. In such position and orientation, quercetin increases the average distance among phospholipids, reducing their mutual attraction and therefore softening the membrane.

References

[1] A. T. Jan, M. R. Kamli, I. Murtaza, J. B. Singh, A. Ali, et Q. M. R. Haq, « Dietary Flavonoid Quercetin and Associated Health Benefits—An Overview », *Food Reviews International*, vol. 26, nº 3,302-317,2010,

[2] Z. Huang, S. Hua, Y. Yang, et J. Fang, « Development and evaluation of lipid nanoparticles for camptothecin delivery: a comparison of solid lipid nanoparticles, nanostructured lipid carriers, and lipid emulsion », *Acta Pharmacol Sin*, vol. 29, nº 9,1094-1102,2008

[3] J. Hao *et al.*, « Encapsulation of the flavonoid quercetin with chitosan-coated nano-liposomes », *LWT - Food Science and Technology*, vol. 85, 37-44, 2017,

[4] J. Azzi, A. Jraij, L. Auezova, S. Fourmentin, et H. Greige-Gerges, « Novel findings for quercetin encapsulation and preservation with cyclodextrins, liposomes, and drug-in-cyclodextrin-in-liposomes », *Food Hydrocolloids*, vol. 81,328-340,2018,

[5] A. C. Anselmo et S. Mitragotri, « Impact of particle elasticity on particle-based drug delivery systems », *Advanced Drug Delivery Reviews*, vol. 108,51-67,2017,

Pseudoquantifying hundreds of proteins in millions of single cells

Etienne Becht[1,2], Evan W. Newell[1], Raphael Gottardo[1] and Mark Headley[1]
[1] Vaccine and Infectious Diseases Division, Fred Hutchinson Cancer Research Center, Seattle, WA, USA.
[2] Institut de Recherches Internationales Servier, Translational Medicine, Suresnes, France

Corresponding author: etienne.becht@servier.com

Reference paper: **Becht *et al.* (2021) High-throughput single-cell quantification of hundreds of proteins using conventional flow cytometry and machine learning. *Science Advances*.** https://doi.org/10.1126/sciadv.abg0505

1. Introduction

Multicolor flow cytometry (FC) is an experimental technique that allows the pseudoquantification of multiple proteins in single cells. FC is commonly used to perform in-depth characterization of tissues with complex cellular compositions. Although FC has been developed decades ago, it currently only allows the quantification of about 20 proteins per experiment. Alternative techniques to FC have been developed such as mass-cytometry, oligonucleotide-barcoded antibodies or single cell mass-spectrometry[1]. These techniques are however more limited in the number of cells that can be profiled, are more expensive and require less common expertise. To overcome these limitations, we have developed *Infinity Flow*, a mixed experimental and computational workflow allowing the quantification of hundreds of proteins across million of single cells.

2. Methods

Experimentally, the sample is divided into $n \approx 300$ subsamples. A set of common *"backbone"* proteins is quantified on all the subsamples, while n distinct *"exploratory"* proteins are each uniquely quantified in one subsample. Traditional techniques used for the analysis of FC data are however unable to incorporate information from the quantification of *exploratory* proteins to inform upon cell phenotypes.

We thus developed a computational workflow tailored for the analysis of these experiments. Its main contribution is to train n multivariate non-linear regression models that each attempt to predict the expression level a single *exploratory* protein by using the expression values of the *backbone* proteins as predictors. These models are then used to predict the expression levels of the n *exploratory* proteins across the n subsamples, enabling the consolidation of the n partially-disjoints datasets representing each subsample into a single cohesive dataset of predicted protein expression values. This workflow additionally performs a data-correction step of non-specific signal, a commonly-known limitation of FC, by linearly-modeling the predicted expression with the predicted expression of a corresponding negative control.

3. Results

This workflow is available as a R Bioconductor package (https://www.bioconductor.org/packages/infinityFlow/). It is designed to be easily applied by users with limited computational experience, using a single public function that applies the whole workflow and generates outputs in the standard FC file format for downstream analyses and visualizations supported by dimensionality reduction.

By using this method, we were able to deeply profile the cellular constituency of the mouse lung, annotating 97% of the cells versus 37.7% for the standard analysis method. We showed that applying clustering or non-linear dimensionality reduction to the predicted data allowed the resolution of additional cellular phenotypes and better delineated distinct cell types. This workflow thus provides more informative data than the *backbone* alone. By profiling lung samples in a melanoma metastasis model, we identified new heterogeneity in phagocytes, enabling a better understanding of the immune response to early lung metastasis.

References

[1] L. Liu, D. Chen, J. Wang, and J. Chen. Advances of Single-Cell Protein Analysis. *Cells*, 9(5), 05 2020.

Session 5, Amphi B

Integrative genomics & evolution

Overcoming small and incomplete datasets limitations in multi-omic integration

Pauline Brochet[1], Edecio Cunha-Neto[2], Christophe Chevillard[1], Lionel Spinelli[1]

[1] INSERM, UMR_1090, Aix Marseille Université, TAGC Theories and Approaches of Genomic Complexity, Institut MarMaRa, Marseille, France.

[2] Laboratory of Immunology, Heart Institute (InCor), University of São Paulo, School of Medicine, São Paulo, Brazil.

Corresponding Author: pauline.brochet@univ-amu.fr

Abstract. *With the progress in large-scale analysis, many single-omics analysis have been conducted. In order to have an overview of the biological process, new methods integrating all these omics together have been developed, called "multi-omics". However, performing this multi-integration may be a challenge. In this study, we performed multi-omics integration on Chagas disease Cardiomyopathy data, containing missing samples and a low number of subjects. Therefore, we proposed some approaches to overcome these limitations.*

Keywords. Multi-omics, imputation, data augmentation

1. Introduction

With the progress made in large-scale analysis [1], many omics analyses have been conducted, bringing significant advances, particularly in the field of medicine [2]. However, these studies each focus on a particular aspect of the biological process, limiting the study of pathology. In recent years, new methods called "multi-omics" have emerged, allowing integration of multiple single-omics datasets. The benefit of multi-omics approaches compared to single-omics has already been demonstrated in many studies, whether it be the discovery of new pathways [3] or biomarkers [4]. Most of the time the single omic analyses were done on different samples inducing missing data for multi omic integration. Excluding samples to get a full set is not appropriate as it will lead to lack of power. Although some multi-omics methods, such as MOFA+ [5], try to deal with this limitation, allowing a small proportion of missing values, another possibility is to impute the missing samples. In fact, numerous imputation methods have already been developed [6] and begin to be adapted for multi-omics dataset [7]. Another limitation in multi-omics analysis is the size of the dataset. Indeed, due to the cost or technical limitations, some analyses are limited by a small cohort [8]. Considering the heterogeneity of omics data, a low number of samples may result in a lack of statistical power, and therefore may impact the relevance of the results [9]. In machine learning classification problems, many methods have been developed to oversampling samples, from bootstrapping [10] (create artificial sample from noising expression of true sample) to methods based on K-Nearest Neighbors [11]. Some of these methods are beginning to be used on multi-omics data [12]. On the other hand, some multi-omics algorithms, like DIABLO [13], try to deal with this by performing a factorization of each omics data, and then integrate those factors all together. In this study, we will evaluate new approaches to overcome small and incomplete datasets limitation using imputation and oversampling methods. We apply these methods on the framework of Chagas disease.

Chagas disease is a neglected disease caused by the protozoan *Trypanosoma cruzi*. The clinical course of the disease comprises an acute phase, mostly asymptomatic, and a chronic phase, where 60% of the patients remain asymptomatic [14]. However, 40% develop symptomatic disease, being 10% megaviscera or neuropathy [15], and 30% Chagas disease cardiomyopathy (CCC) with varying degrees of severity including refractory heart failure [14]. This cardiomyopathy is the main cause of deaths from Chagas disease itself, and is one of the most lethal cardiomyopathies [16]. The fact that the biological processes leading to CCC are not yet well understood has impaired the development of efficient therapeutic strategies. In order to understand the pathogenic process, we have now set up a multi-omics approach, including transcriptomic, methylomic, miRNomic and mitochondria dataset. Since we work on human heart tissue, we are confronted with both the problem of missing samples and small size in our cohort. In this study, we performed multi-omics analysis, with a special focus on pre-processing methods adapted to our data.

2. Methods

2.1. Patients and Myocardial Tissue Collection

Human left ventricular free wall heart tissue samples were obtained from patients with end-stage heart failure CCC at the time of heart transplantation (n=14). CCC patients underwent a serological diagnosis of T. cruzi infection and standard electrocardiography and echocardiography, and tissues were subject to histopathological assessment as previously described. Biopsies from controls (n=7) were obtained from healthy hearts of organ donors having no suitable recipient.

2.2. Single omics preprocessing

Four different studies were performed from human heart tissue: bulk RNA-sequencing, miRNA sequencing, Illumina EPIC methylation and mitochondria variant calling. Each dataset was aligned on the GRCh37 (hg19) human reference genome from Ensembl, and all classical pre-processing steps have been applied as described in previous study [17,18]. Due to the specificities of each omics, the feature (genes, CpG site…) values are not comparable between them: a gene may have an expression value from 0 to +∞, while a CpG ranges from 0 to 1. All omics values have therefore been centered and scaled to be comparable. Each single omics dataset also has a different number of features: 57,905 genes, 850,000 CpGs, 758 miRNAs and 12429 mitochondria variants.

2.3. Missing sample imputation

The multi-omics dataset includes a total of 21 subjects (7 controls, 14 cases). A total of 6 samples are missing in RNAseq data, 2 in methylation, 5 in miRNAseq and 1 in mitochondria (table 1). In order to complete the multi-omics dataset, missing values for samples were imputed. This imputation was performed using the KNN imputer function from scikit-learn [19]. This method imputes the missing values according to the K-nearest neighbor. However, multi-omics integration is performed in multiple datasets, so the first neighbor in one omic is not necessarily the same in another omic. A custom matrix distance for the multi-omics dataset was provided to the KNN imputer function, taking into account for each omic: the distance between all samples, the number of features and the number of missing samples. The imputation of controls and cases was done independently. For each phenotype, the following steps have been performed (as described in figure 1) :

Step 1. For each omic M and for each couple of samples (i,j) in M, Euclidean distance metrics has been computed.

Step 2. For each distance matrix obtained from step 1, each distance value between a couple of samples (i,j) was divided by the number of features N in omic M. Since the number of features is very different from one omic to another, this step allows to normalize the distance between two samples in each omic by the number of features.

Step 3. An average distance matrix has been computed. The normalization term of the mean corresponds to the number of available samples through omics.

Step 4. The missing samples imputation is performed with the KNN imputer function of sklearn, using the custom distance matrix made previously.

By this way, both differences in omic size and number of missing samples are normalized in the final distance matrix, and all omics have the same contribution in data imputation.

2.4. Data augmentation

Considering the small number of subjects available (n=21), a statistical power analysis with the MultiPower [20] tool has been made. The resulting low-power level (see results) suggests the need to increase the subject number. Therefore, synthetic subjects were generated. Assuming that the data space of each phenotype is convex, we build a new synthetic subject as a linear combination of p samples with the same phenotype, each with a random weight (with sum of weights equal to 1). Therefore, if the number n of synthetic samples tends to infinity, it is expected that the synthetic samples are uniformly distributed in the data space. However, the features in a specific assay being constrained (dependencies between features due to biological functions and processes), it is expected that the samples in a feature space are distributed on a hyperspace of reduced dimension, often described as a manifold. Since it cannot be assumed that this manifold is a linear subspace, a linear combination of samples may violate the manifold constraints, producing a synthetic sample that is not on the manifold and not biologically relevant. So, it is important to control the distance of the produced synthetic samples to the manifold. Since it is not possible to estimate the manifold with few samples, the best guess is to ensure that the synthetic data will not fall far from existing samples, hence limiting the violation of the biological constraints. The data augmentation method applied in this study is based on this reasoning. Assuming that the space of each class is convex, n samples of the same class $P_1, P_2, \ldots, P_n$ are randomly selected. The new synthetic sample P' will be placed at the barycenter of those points, a random weight, between 0 and 1 being added to each sample. The sum of all weight must be equal to 1 and the first weight, called α, was chosen to be between 0.8 and 1. Thus, synthetic samples are linear combinations of n raw samples of the same class, but each time, one random sample is favored compared to the others.

2.5. Feature selection

The total multi-omic dataset contains more to 900.000 features, for 21 raw samples. Even with the amplification, the imbalance between the number of features (n) and samples (p) (n << p) is a well-known problem in large-scale data analysis. The main risk is overfitting as too many features are provided. Therefore, a feature selection must be done before multi-omics integration. As all the omics data are continuous, a Wilcoxon test has to be applied between control and case samples. Each feature has been sorted according to their significance, and the top-1000 features have been selected for next analysis.

2.6. Multi-omics integration

The multi-omics integration has been made with the DIABLO method from the MixOmics R package. Each omic has the same weight in the model, and the DIABLO correlation cut-off has been set up to 0.8. Biological Process Gene Ontology enrichment analysis has been made with the ClueGo Cystoscape plugin on the selected features obtained with DIABLO.

2.7. Workflow optimization and validation

In order to adjust the parameters of each step to our data, the following method was implemented: for each possible parameter (in a predefined list), the complete analysis was performed 10 times (runs). Considering that the data augmentation is done in a random way (random selection of the raw subjects and randomly assigned weights), each run provides different synthetic samples, and so a different final matrix. Finally, the results of each run were compared together to select the optimal parameters. Firstly, the intersection of the results between the 10 runs was made. Then, for each run, the percentage of common features between the run and this intersection was computed. By this way, the robustness of the analysis with the selected parameters was evaluated. On the other hand, a validation step was performed. The principle used here was quite similar to a leave-one-out cross validation. As before, 10 similar analyses were made, but each time, an existing sample in one omic was randomly removed. The results obtained were compared to the raw results, without removing this sample. Thus, the accuracy of the imputation, supported by the augmentation of samples around the imputed samples, could be measured. Two metrics have been used to compare those results: the Jaccard index (intersection of two datasets divided by their union) and the minimal overlap (intersection of two datasets divided by the smallest dataset), allowing an overview of the similarity of those two approaches. The same analysis was made by augmented data with bootstraping, giving a random noise (from 0 to 1) to each value of a subject, and the results between the two data augmentation methods were compared.

3. Results

3.1. Single-omic preprocessing

The current analysis is focused on the integration of 4 single-omics dataset: bulk RNAseq, methylation, miRNAseq and mitochondrial sequencing. Those 4 datasets contain continuous features, ranging from 0 to +∞ (RNAseq and miRNAseq) or from 0 to 1 (methylation and mitochondria). The feature number varies a lot between dataset. Indeed, after quality control steps, a total of 43,850 genes are considered in RNAseq, 710,846 CpG in methylation, 652 miRNA in miRNAseq and 12,325 variants in mitochondria (table 1).

	Nb. of features	CTRL							CCC													
		1	*2*	*3*	*4*	*5*	*6*	*7*	*1*	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*	*10*	*11*	*12*	*13*	*14*
RNAseq	43,850	x	x	x	x	x	x	x		x	x		x		x	x		x		x		x
Methylation	710,846	x	x	x	x		x	x		x	x	x	x	x	x	x	x	x	x	x	x	x
miRNAseq	652		x	x	x	x	x	x	x	x	x		x	x	x	x		x		x		x
Mitochondria	12,325	x	x	x	x		x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x

Table 1. Number of available features and samples in each omics

3.2. Estimation of optimal number of samples

Giving some information on single-omic dataset, such as standard deviation, number of features, expected proportion of differentially expressed features…, the MultiPower tool can estimate the optimal number of samples sufficient to have a high enough statistical power. According to MultiPower results, with the initial subjects (n = 21, 7 controls and 14 cases), only the RNAseq and methylation datasets are good enough to have a correct statistical power (approximately 0.75). However, miRNA and mitochondria are not able to provide powerful analyses (power < 0.3). To have a sufficient statistical power (up to 0.8), the number of samples should be increased in each omic, more or less depending on the omics, by a factor of 1.5 (methylation) to 5 (miRNA). In conclusion, based on these estimates, the optimal number of samples per phenotype should be about 100 (figure 1).

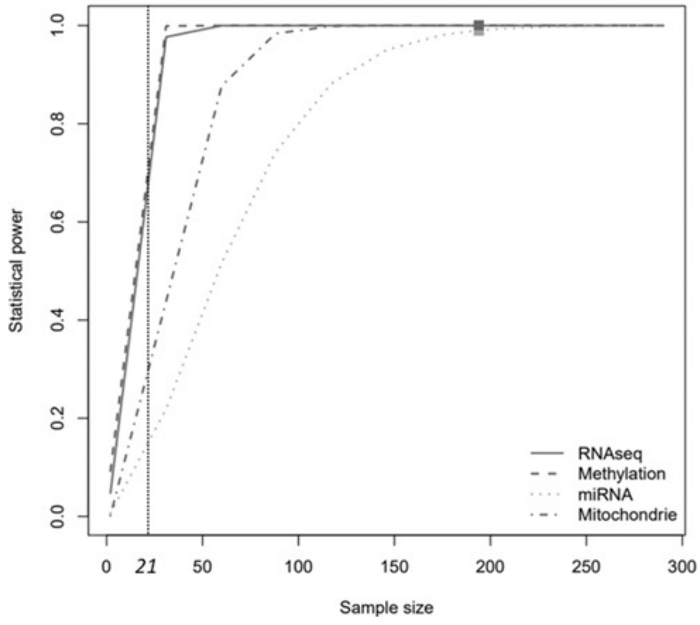


Figure 1. Statisical power analysis performed with MultiPower tool. An estimation of the statistical power is representing according to the number of samples for each omics.

3.3. Optimization of data augmentation parameters

Parameters selection was performed using the protocol described in methods. Concerning the α parameters chosen, the top N selected features remain almost the same. The choice of the alpha parameter does not seem to have a great consequence on the selection of the features. However, as we stated in the method section, in order to avoid a discrepancy in the constraints between the omics features, we need to keep the synthetic subjects close to real subjects but still adding some variability. To do so, we decide to set the range of α value to [0.8;1]. Another parameter to take into account when augmenting the data is the number of subjects S to reach. According to the power analysis previously performed, a final number of 100 subjects per phenotype seems optimal. However, this calculation is based on estimates only. The same approach as for the choice of the range was therefore used. The number of subjects has an important impact on the selected features. Indeed, a low number of subjects lead to important variabilities in the choice of features between runs. Conversely, a large number of subjects reduces these variabilities, and increases the percentage of common features found for each omic. In conclusion, increasing the subject size seems to reduce the relative noise due to a low statistical power, and amplifies the signal relative to the phenotypes. Therefore, the number of subject by group has been set up to 200.

3.4. Feature selection

Considering the higher value of features compared to the number of subjects, a feature selection step was applied (see methods). With the same analysis as before, different results are obtained depending on the omics. Concerning mitochondrial data, almost 90% of features were retrieved, taking the top 1000 variants. Since miRNAs contains less than 1000 features, and other omics show a large percentage of features in common beyond the top 1000 features, the top-1000 features were selected in each omic.

3.5. Multi-omic integration

Final multi-omics integration has been performed with all preprocessing steps described above. After multi-omics integration, 2048 features were selected. The clustering of the DIABLO results interestingly shows a good mixing of the various selected omics features, indicating a strong interplay between omics. The Gene Ontology enrichment analysis shows those features are involved in 4 main biological processes: immune response, nervous system, ion concentration and G-protein receptors. Even if those pathways seem very different, some features make it possible to link them, such as IFNG, which is involved in neuron generation and the differentiation of T cells into Th1 or Th2. Those features, or features interacted with them, may be key elements in the development of CCC.

3.6. Workflow performance estimations

In order to ensure the robustness of our pipeline, an analysis similar to cross-validation was implemented. 10 runs were performed, each time randomly removing a real sample in a given omic.In a first step, this approach allowed us to have an estimate of the imputation performance. Indeed, the imputed samples are not identical to the true samples, with an Euclidean distance from 0.5 to 8e+06. Omitting a true sample may change the imputation of other missing samples, since our imputation method is based on the k nearest neighbors. Removing one will therefore necessarily bias the imputation of these samples. However, excluding one sample does not significantly affect the results and the imputed samples remain very strongly correlated with their original samples (from 0.96 to 0.99). To confirm this hypothesis, the final results obtained after multi-omic integration were compared. Regardless of the samples removed and the omics affected, a large part of the features selected after multi-omic integration are identical (80 to 95%). These results tend to confirm the high robustness of our protocol (figure 2A). On the other hand, using bootstraping methods, each run provides different results, with a small percent of common features selected by the multi-omics integration (37 to 40%), illustrating the importance of adding some variability in the synthetic samples (figure 2B).

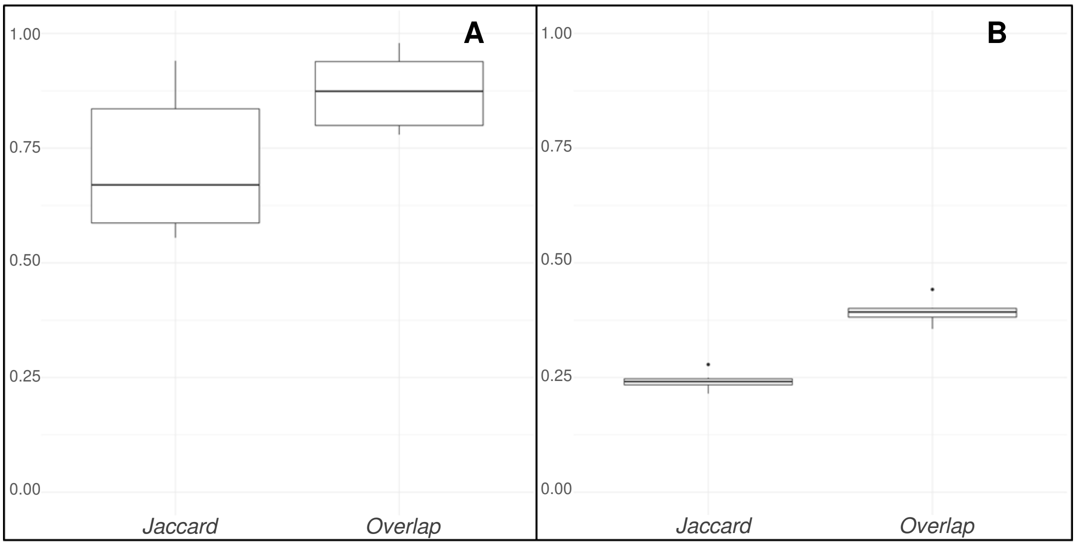


Figure 2. Distribution of Jaccard index and minimal overlap through the 10 analysis. A) Data augmentation performed with the barycenter method. B) Data augmentation performed with the bootstraping method.

4. Discussion

Multi-omic integration is a approach of growing interest to identify complex biological processes. These approaches are particularly interesting for the understanding and treatment of certain diseases. In this study, we focused on Chagas disease. Due to the pathology, the experiments were performed on human heart tissues, limiting the number of available subjects and thus the statistical power of the study. Moreover, as not all the experiments were performed at the same time, the same subjects were analyzed with all the same omics assays, generating missing values in the dataset. In this article, we presented methods to counteract these problems.

Missing values is a well-known problem in statistical analysis. One of the most used methods to overcome this issue is to impute a missing value as the mean of all the values for the same variable. However, this method is limited when all the values have a great variance, and it cannot be applied to more than one missing value. Other methods, based on dimensionality-reduction methods (like PCA), have been developed. The principle is to impute the missing values with an initial value, such as the mean of the variable, or a value from other samples. Then, an algorithm like PCA or MCA is applied on the dataset to understand the relations between each sample, and use them to impute the missing values. Based on the same principles, some methods perform N multiple imputation (MI) by randomly selecting different initial values, and use those multiple final results to correctly impute the missing values. That’s the case of the MI-MFA [6] tool, an algorithm developed to impute missing values in a multi-omics dataset. In that case, the dimensionality-reduction step was made with a MFA, and all the resulting configurations were resumed in one with the STATIS [21] method. However, all these methods don't take into account the specificity of each omics dataset. On the other hand, some imputation approaches are based on the k-nearest neighbors algorithm [22]. In that case, a missing value is not imputed thanks to all the known values, but only according to the closest other values. By creating a custom matrix distance, each omics has the same weight in the KNN selection, ensuring that one omic doesn’t drive the analysis. In our study, we used such an approach but employing a customized distance matrix, able to take into account specificity on missing data and omics features number. As result, the leave-one-out cross validation has shown that the imputed samples are not perfectly identical to the original samples. Nevertheless, they remain strongly correlated to the values they should have had, and a large part of the features selected after multi-omics integration is always found, ensuring the robustness of our approach.

Considering the reduced statistical power due to the small size of our dataset, we chose to increase the number of subjects in our data by creating synthetic samples. One of the most popular data augmentation algorithms is SMOTE [11]. In fact, this method has been developed to oversample a class when this class is in minority respect to others. In our situation, we employ this kind of method to enlarge the dataset. SMOTE creates a synthetic sample as a linear combination of one random selected sample and one of its k-nearest neighbors, given random weight to those two samples.

However, SMOTE has some disadvantages: it may decrease data variability, oversample noisy samples or increase the overlap between different classes around the class boundaries [23]. But the more important limitation of SMOTE is that samples are no longer independent. Indeed, because one synthetic sample is a linear combination of only two samples, the independence hypothesis, required for many methods is no longer respected. Many variants of SMOTE have been developed through the years [24]. Among them, KNNOR [25] is not limited to only two samples for the creation of synthetic samples. In fact, KNNOR acts like SMOTE, but repeats the SMOTE process several times to finally obtain a synthetic subject that is a linear combination of several subjects. By this way the final sample depends not on 2 real subjects, but on several real subjects. However, considering that subjects belonging to different classes can be mixed together (for example, one cluster of control in the middle of cases), KNNOR considers the last nearest neighbors as outliers and removes it. In biological dataset, it is impossible to be sure that a sample far from the others (in terms of Euclidean distance) is an outlier, and does not carry relevant information. Moreover, in our multi-omic dataset, the data space of each phenotype seems to be convex. So, removing one sample will only reduce the data variability which is not desirable. SMOTE and all its derived methods are based on the selection of the k-nearest neighbors. As it was discussed before, the k-nearest neighbors for one omics may not be the same for another omics. Because all our omics present a huge disequilibrium in the feature number, the neighbors selection may be influenced by one dataset, inducing a bias in the data augmentation. The same approaches used in imputation could be applied here, but considering the convexity of our spaces, this approach does not seem essential. Moreover, the KNN in linear space aren't necessarily the same nearest neighbors in the manifold carrying the subject in the feature space. For all these reasons, we choose to use another method, based on the barycenter computation. By adding a dominant weight to one sample compared to others, we ensure to stay close to the subjects carrying manifolds, and so to samples that can be biologically realistic. In order to evaluate the quality of our approach, we applied a leave-one-out cross-validation method. Removing one subject at a time, we repeatedly applied our pipeline with missing data imputation, data augmentation and feature selection (see methods) several times. Each try was subject to multi-omics analysis using DIABLO. The results show a large stability of the multi-omic analysis, confirming the robustness of our approach.

Applied to a complete dataset, our pipeline provided interesting first results, in line with the current knowledge of CCC, but also less-expected process. For example, neurological disorders have previously been identified in CCC, but their importance in disease development has not been demonstrated yet. IFNG, one of the key elements in CCC, has also been raised by our analysis in interplay between both immune response and nervous system processes. Some roles of IFNG in the nervous system have already been described [26], but never in CCC. Moreover, G protein-coupled receptors have also been identified. Some autoantibodies have already been found against those receptors in both Chagas disease cardiomyopathy and megacolon chronic forms [27]. Considering the enrichment in features involved in T-cell receptor processes, all this information taking together provides additional insight into the pathogenic process leading to CCC.

Acknowledgements

Centre de Calcul Intensif d'Aix-Marseille is acknowledged for granting access to its high performance computing resources.

References

1. S. Goodwin, J. D. McPherson, et W. R. McCombie, « Coming of age: ten years of next-generation sequencing technologies », *Nat. Rev. Genet.*, vol. 17, nº 6, p. 333-351, mai 2016, doi: 10.1038/nrg.2016.49.
2. E. A. Ashley, « Towards precision medicine », *Nat. Rev. Genet.*, vol. 17, nº 9, Art. nº 9, sept. 2016, doi: 10.1038/nrg.2016.86.
3. P. M. Quirós *et al.*, « Multi-omics analysis identifies ATF4 as a key regulator of the mitochondrial stress response in mammals », *J. Cell Biol.*, vol. 216, nº 7, p. 2027-2045, juill. 2017, doi: 10.1083/jcb.201702058.
4. D. Chantzichristos *et al.*, « Identification of human glucocorticoid response markers using integrated multi-omic analysis from a randomized crossover trial », *eLife*, vol. 10, p. e62236, avr. 2021, doi: 10.7554/eLife.62236.

5. R. Argelaguet *et al.*, « MOFA+: a statistical framework for comprehensive integration of multi-modal single-cell data », *Genome Biol.*, vol. 21, n° 1, p. 111, mai 2020, doi: 10.1186/s13059-020-02015-1.
6. F. Husson et J. Josse, « Handling missing values in multiple factor analysis », *Food Qual. Prefer.*, vol. 30, n° 2, p. 77-85, déc. 2013, doi: 10.1016/j.foodqual.2013.04.013.
7. V. Voillet, P. Besse, L. Liaubet, M. San Cristobal, et I. González, « Handling missing rows in multi-omics data integration: multiple imputation in multiple factor analysis framework », *BMC Bioinformatics*, vol. 17, n° 1, p. 402, oct. 2016, doi: 10.1186/s12859-016-1273-5.
8. C. P. Shannon *et al.*, « Multi-Omic Data Integration Allows Baseline Immune Signatures to Predict Hepatitis B Vaccine Response in a Small Cohort », *Front. Immunol.*, vol. 11, p. 578801, 2020, doi: 10.3389/fimmu.2020.578801.
9. K. S. Button *et al.*, « Power failure: why small sample size undermines the reliability of neuroscience », *Nat. Rev. Neurosci.*, vol. 14, n° 5, p. 365-376, mai 2013, doi: 10.1038/nrn3475.
10. G. Bao *et al.*, « Data Augmentation for EEG-Based Emotion Recognition Using Generative Adversarial Networks », *Front. Comput. Neurosci.*, vol. 15, 2021, Consulté le: 9 mars 2022. [En ligne]. Disponible sur: https://www.frontiersin.org/article/10.3389/fncom.2021.723843
11. Nitesh V. Chawla, Kevin W. Bowyer, Lawrence O. Hall, and W. Philip Kegelmeyer. 2002. SMOTE: synthetic minority over-sampling technique. J. Artif. Int. Res. 16, 1 (January 2002), 321–357.
12. M. T. Hira, M. A. Razzaque, C. Angione, J. Scrivens, S. Sawan, et M. Sarker, « Integrated multi-omics analysis of ovarian cancer using variational autoencoders », *Sci. Rep.*, vol. 11, n° 1, Art. n° 1, mars 2021, doi: 10.1038/s41598-021-85285-4.
13. A.Singh *et al.*, « DIABLO: an integrative approach for identifying key molecular drivers from multi-omics assays », *Bioinforma. Oxf. Engl.*, vol. 35, n° 17, p. 3055-3062, 01 2019, doi: 10.1093/bioinformatics/bty1054.
14. J. A. Pérez-Molina et I. Molina, « Chagas disease », *The Lancet*, vol. 391, n° 10115, p. 82-94, janv. 2018, doi: 10.1016/S0140-6736(17)31612-4.
15. M. O. Py, « Neurologic manifestations of Chagas disease », *Curr. Neurol. Neurosci. Rep.*, vol. 11, n° 6, p. 536-542, déc. 2011, doi: 10.1007/s11910-011-0225-8.
16. P. J. Hotez *et al.*, « An Unfolding Tragedy of Chagas Disease in North America », *PLoS Negl. Trop. Dis.*, vol. 7, n° 10, oct. 2013, doi: 10.1371/journal.pntd.0002300.
17. L. Laugier *et al.*, « miRNAs may play a major role in the control of gene expression in key pathobiological processes in Chagas disease cardiomyopathy », *PLoS Negl. Trop. Dis.*, vol. 14, n° 12, p. e0008889, déc. 2020, doi: 10.1371/journal.pntd.0008889.
18. P. Brochet *et al.*, « Specific methylation marks in promoter regions are associated to the pathogenic process of Chronic Chagas disease Cardiomyopathy by modifying transcription factor binding patterns ». medRxiv, p. 2021.12.20.21267972, 21 décembre 2021. doi: 10.1101/2021.12.20.21267972.
19. F. Pedregosa *et al.*, « Scikit-learn: Machine Learning in Python », *J. Mach. Learn. Res.*, vol. 12, janv. 2012.
20. Tarazona, S., Balzano-Nogueira, L., Gómez-Cabrero, D. et al. Harmonization of quality metrics and power calculation in multi-omic studies. Nat Commun 11,3092 (2020).https://doi.org/10.1038/s41467-020-16937-8
21. C. Lavit, Y. Escoufier, R. Sabatier, et P. Traissac, « The ACT (STATIS method) », *Comput. Stat. Data Anal.*, vol. 18, n° 1, p. 97-119, août 1994, doi: 10.1016/0167-9473(94)90134-1.
22. K. M. Fouad, M. M. Ismail, A. T. Azar, et M. M. Arafa, « Advanced methods for missing values imputation based on similarity learning », *PeerJ Comput. Sci.*, vol. 7, p. e619, 2021, doi: 10.7717/peerj-cs.619.
23. P. Soltanzadeh et M. Hashemzadeh, « RCSMOTE: Range-Controlled synthetic minority over-sampling technique for handling the class imbalance problem », *Inf. Sci.*, vol. 542, p. 92-111, janv. 2021, doi: 10.1016/j.ins.2020.07.014.
24. G. Kovács, « An empirical comparison and evaluation of minority oversampling techniques on a large number of imbalanced datasets », *Appl. Soft Comput.*, vol. 83, p. 105662, oct. 2019, doi: 10.1016/j.asoc.2019.105662.
25. A.Islam, S. B. Belhaouari, A. U. Rehman, et H. Bensmail, « KNNOR: An oversampling technique for imbalanced datasets », *Appl. Soft Comput.*, vol. 115, p. 108288, janv. 2022, doi: 10.1016/j.asoc.2021.108288.
26. P. A. Ottum, G. Arellano, L. I. Reyes, M. Iruretagoyena, et R. Naves, « Opposing Roles of Interferon-Gamma on Cells of the Central Nervous System in Autoimmune Neuroinflammation », *Front. Immunol.*, vol. 6, 2015, Consulté le: 16 mars 2022. [En ligne]. Disponible sur: https://www.frontiersin.org/article/10.3389/fimmu.2015.00539
27. S. G. Muñoz Saravia *et al.*, « Distinct Patterns of Autoantibodies Against G-Protein–Coupled Receptors in Chagas' Cardiomyopathy and Megacolon: Their Potential Impact for Early Risk Assessment in Asymptomatic Chagas' Patients », *J. Am. Coll. Cardiol.*, vol. 56, n° 6, p. 526-527, août 2010, doi: 10.1016/j.jacc.2010.02.057.

Interaction between gene expression and chromatin environment: application to the study of H3K79me2 and its methyltransferase DOT1L during spermatogenesis

Manon Coulée[1], Clara Gobé[1], Mélina Blanco[1], Clémentine Lapoujade[2], Romain Daveau[3], Pierre Fouchet[3], Guillaume Meurice[3], Julie Cocquet[1*], Laila El Khattabi[1,4*]

[1] Institut Cochin, INSERM U1016, CNRS UMR8104, Université Paris Cité, Paris, France
[2] Université Paris Cité and Université Paris-Saclay, iRCM/IBFJ CEA, UMR Stabilité Génétique Cellules Souches et Radiations, Laboratoire des Cellules Souches Germinales, Fontenay-aux-Roses, France
[3] MOABI-APHP (Bioinformatics platform), Paris, France
[4] APHP centre-université Paris Cité, service de génétique, Paris, France

[*]Co-last,co-corresponding authors : julie.cocquet@inserm.fr & laila.el-khattabi@aphp.fr

This work was supported by the Agence Nationale de la Recherche (ANR-17-CE12-0004-01 to J.C.), the Fondation pour la Recherche Médicale (SPF201909009274 to C.G.) and the Institut Cochin (PIC to L.E.K).

Abstract

Spermatogenesis, the biological process during which male germ cells differentiate into spermatozoa, is very dynamic in terms of gene expression and chromatin remodeling. During the last phase of spermatogenesis, haploid cells called spermatids undergo profound morphological and functional changes, in particular an extreme remodeling of their chromatin, leading to the replacement of most histones by protamines. DOT1L is the only H3K79me2 methyltransferase and is highly expressed in spermatids. H3K79me2 level is also very high just before histone removal. To understand the role of Dot1l and H3K79me2 during sperm differentiation, we have produced a mouse model in which Dot1l gene is conditionally knocked-out (KO) in male germ cells. In this study, we have analyzed H3K79me2 ChIPseq and RNAseq data from Dot1l-KO cells. We have also used ChromHMM tool to integrate these data with chromatin environment annotations. First, we observed that, as in other cell types, H3K79me2 mark has a sharp signal along the gene body and is associated with high levels of gene expression in male germ cells. We also found that H3K79me2 is enriched at enhancers and that enhancers marked with H3K79me2 are associated with a higher expression level of nearby genes. Surprisingly, we found that Dot1l-KO produces more upregulated genes than downregulated genes; this deregulation is not strictly correlated with H3K79me2 enrichment but ChromHMM annotation indicates that deregulated genes are associated with bivalent and repressive states. These states are characterized by the presence of a repressive histone mark, H3K27me3, suggesting a link between DOT1L and H3K27me3, independent of H3K79 methylation.

Keywords

RNAseq ; ChIPseq ; ChromHMM ; H3K79me2 ; spermatogenesis

1 Introduction

Spermatogenesis is the biological process during which male germ cells differentiate into spermatozoa. It is a very dynamic differentiation process in terms of gene expression and chromatin remodeling [1-4]. Spermatogenesis can be divided in 3 phases: (i) the proliferation of spermatogonia (which have a 2N DNA content) which then differentiate into primary spermatocytes and replicate their DNA (4N); (ii) meiosis which consists in two divisions producing secondary spermatocytes (2N) then round spermatids (N); (iii) the last phase, during which spermatids undergo profound morphological and functional changes to become spermatozoa. This last step, called spermiogenesis, is characterized by the replacement of almost all histones by protamines allowing a high level of chromatin compaction that is essential to maintain the integrity of the genetic material [5]. A crucial step of this global chromatin remodeling process is the hyperacetylation of histone 4 which occurs in elongated spermatids just before histone removal. Multiple studies have indeed shown that H4 hyperacetylation (H4ac) is indispensable to the production of spermatozoa with a normal chromatin packaging [6-9]. Interestingly, high levels of H3K79me2 (dimethylation of the lysine at position 79 of histone H3) have been observed at the same time than H4ac and could regulate the remodeling of spermatid chromatin at the end of spermiogenesis [10].

Dot1l (Disruptor of telomeric silencing 1-like) encodes the only methyl transferase able to methylated H3K79 [11] in its 3 forms: mono, di and tri-methylation. All three forms have been associated with gene expression [12]. *Dot1l* has been shown to play a role in the regulation of gene expression and in embryonic development [13]. During spermatogenesis, it is highly expressed in round and elongated spermatids [14]. To understand the role of *Dot1l* and H3K79me2 during sperm differentiation we have produced a mouse model in which *Dot1l* gene is knocked-out in male germ cells (i.e. conditional deletion of exon 2 of the gene). In brief, *Dot1l*-KO males are subfertile with an abnormal histone-protamine transition leading to malformed spermatozoa with abnormal nuclear compaction (Blanco et al, submitted). To characterize the molecular changes associated with this phenotype, we have generated RNAseq and ChIPseq data.

The objective of the present study is to characterize H3K79me2 mark and its dynamics throughout spermatogenesis and to correlate it with transcriptional analyses (RNAseq) obtained from *Dot1l*-KO male germ cells. For this, we used the computational algorithm ChromHMM [15,16] to annotate H3K79me2 mark and to link *Dot1l*-KO deregulated genes to their chromatin environment. ChromHMM uses a combination of chromatin marks ChIPseq datasets to assign the most probable chromatin state to each genomic segment. A chromatin state is characterized by one or more histone marks and is annotated according to its link with gene expression such as a promoter, enhancer or repressed region.

As in other studies, we focused on H3K79me2 as it correlates the best with gene expression. In fact, all three forms have the same trend but H3K79me1 is very broad and H3K79me3 level is limited [17]. Our study shows that, in male germ cells, H3K79me2 mark is enriched at the gene body and, at enhancers, and that, in both cases, its presence correlates with a higher level of expression. The gene deregulation observed in *Dot1l*-KO is not directly related to H3K79me2 but to the gene chromatin environment, as it correlates with bivalent and repressor chromatin states.

2 Materials et Methods

2.1 Generation of RNAseq and ChIPseq data

Transcriptome analyses were performed on male germ cells from genetically modified mice (« knock-out », KO mice) and control mice (CTL mice). Cells were obtained from several spermatogenesis stages: undifferentiated spermatogonia (Kit-), differentiated spermatogonia (Kit+), primary spermatocytes (SC), secondary spermatocytes (SCII) and round spermatids (RS) (Fig. 1). The different cell types were isolated by flow cytometry sorting according to their size and DNA content. Total RNA was extracted from 1x Kit-, 1x Kit+, 1x SC, 2x SCII, and 4x RS, to correct for genomic content, using the RNAqueous Ambion kit (Thermofisher) following the manufacturer's protocol. Two replicates were generated for each genotype for spermatocyte stages and three replicates were generated for spermatogonia and spermatid stages.

ChIPseq analyses were performed for the following three cell stages: primary spermatocytes, round spermatids and elongated spermatids (ES) obtained by elutriation (Fig. 1). Each of these three stages is represented by two replicates immunoprecipitated with an anti-H3K79me2 antibody (Diagenode C15410051, Lot A1193D) and a sample corresponding to the DNA before immunoprecipitation (input file). Analyses were performed on WT germ cells. ChIP-qPCR and mass spectrometry quantifications confirmed a strong decrease in H3K79me2 mark in KO cells (Blanco et al, submitted).

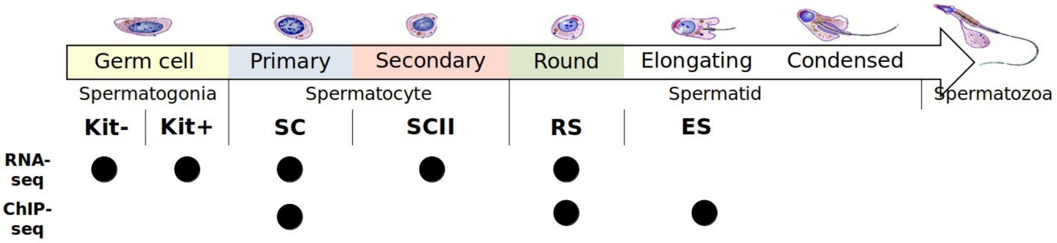


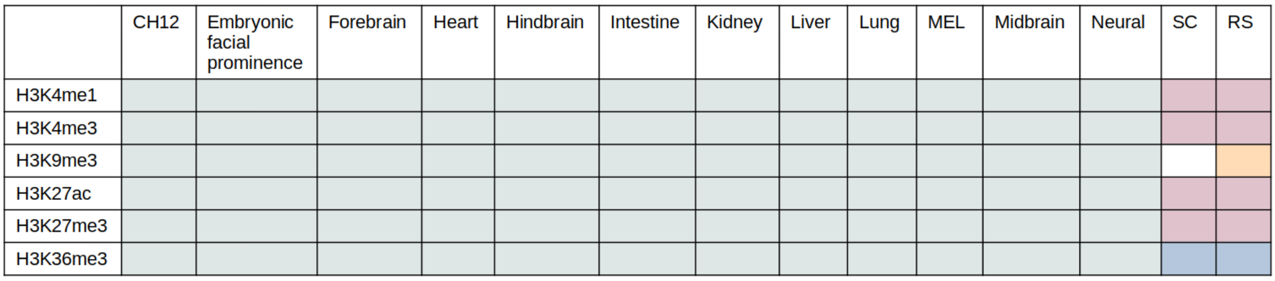
Fig. 1. Cell stages used in transcriptomic and ChIPseq analyses

2.2 RNAseq and ChIPseq analyses

Snakemake was used for RNAseq and ChIPseq data [18] (v. 3.9.0). Adaptors were deleted and reads with poor quality (quality < 20) were filtered with BBduck from BBTools [19]. For RNAseq, the alignment was performed on mouse genome build mm10 using STAR [20](v. 2.7.2d). For each sample, the number of aligned reads was converted in cpm, the number of read per genes was normalized by the sum of the number of reads for all genes. Low expressed genes with an expression < 1cpm in minimum 2 samples were excluded. Differential expression analysis was carried out using DESeq2 [21] and edgeR [22] packages. We designed a model in function genotype, cell type and interaction between cell type and genotype. 20452 genes were defined and divided in 3 categories according to their level of deregulation between KO and control: upregulated genes, with a log(KO/control) > log(1.5) (p-value<5%); downregulated genes, with a log(KO/control) <-log(1.5) (p-value < 5%); all other genes were considered to be "not deregulated".

For ChIPseq, the alignment was performed on mouse genome build mm10 with bowtie2 [23] (v. 2.3.0). Peak calling was performed using MACS2 [24] with --broad parameters. Annotation of peaks were performed using ChIPseeker [25] package with 1.5kb around TSS, and divided in 4 categories: upstream (promoter's annotations), intragenic (exon, intro and UTR), downstream and distal intergenic.

ChromHMM was used with default parameters to predict the chromatin state associated with each genomic region (parameter bin size = 200bp). We created a model of 18 chromatin states from 6 epigenetic marks (H3K4me1, H3K4me3, H3K27ac, H3K27me3, H3K36me3 H3K9me3) originating from 14 different cell types/stages (Fig. 2), retrieved from the literature and the mouse ENCODE project [26-29]



	CH12	Embryonic facial prominence	Forebrain	Heart	Hindbrain	Intestine	Kidney	Liver	Lung	MEL	Midbrain	Neural	SC	RS
H3K4me1	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟥	🟥
H3K4me3	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟥	🟥
H3K9me3	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩		🟧
H3K27ac	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟥	🟥
H3K27me3	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟥	🟥
H3K36me3	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟦	🟦

Fig. 2. Data used to develop ChromHMM model. In green, ENCODE project data; pink, Hammoud *et al* data; orange, Bryant *et al* data; blue, Zuo *et al* data

2.3 Association between enhancer and gene promoter

Among the chromatin states determined with ChromHMM, 4 are characterized as different enhancer states: active enhancers (EnhA), genic enhancers (EnhG), weak enhancers (EnhW) and bivalent enhancers (EnhBiv) Each genomic region characterized as an enhancer was compared with the peak calling values for H3K79me2. We obtained 8 enhancer categories enriched or not in H3K79me2 (EnhG-H3K79me2, EnhG-notH3K79me2, EnhA-H3K79me2, EnhA-notH3K79me2, EnhW-H3K79me2, EnhW-notH3K79me2, EnhBiv-H3K79me2, EnhBiv-notH3K79me2). For each category, we associated to each genomic region, the genes associated to this enhancer. For this purpose, we defined as associated genes, all genes having their TSS within a 500kb interval with respect to the enhancer (enhancer±500kb). For each gene, we associated its expression value (CPM) obtained from CTL RNAseq data.

2.4 Chromatin environment of differentially expressed genes

Each gene identified in our RNAseq analysis, was assigned to its predominant chromatin state. To do this, we estimated 3kb around the gene TSS as the chromatin environment of the gene. We calculated the coverage in base-pair of each chromatin state of the environment to obtain a matrix. A reduction of the model was performed using the umapr package [30] with these following parameters: metric euclidean, number of neighbors 32 and minimal distance 0,99 to obtain a 2-dimensional model clustered according to the predominant state of the gene environment.

3 Results and discussion

3.1 Characterization of H3K79me2 during spermatogenesis

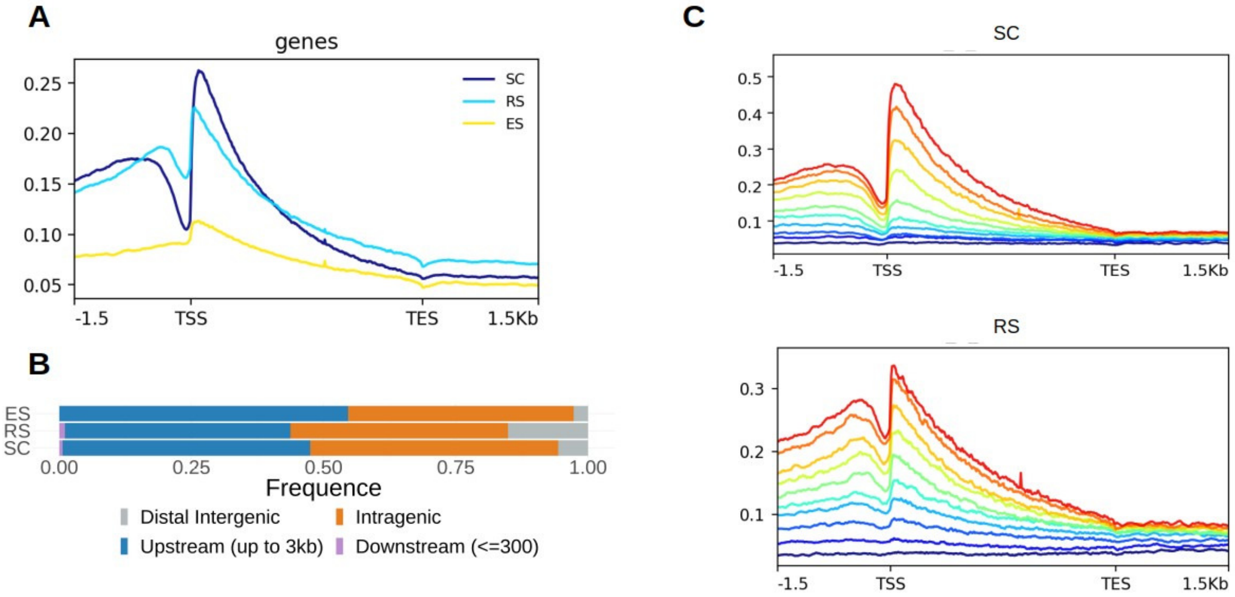


Fig. 3. Characterization of H3K79me2 histone mark in male germ cells by ChIPseq. **A.** H3K79me2 enrichment profile for SC, RS and ES. B**.** Annotation of the peaks identified by the peak calling. **C.** H3K79me2 enrichment according to gene expression level in SC and RS (blue: low expressed genes, red : high expressed genes)

First, to characterize H3K79me2 during spermatogenesis, we performed ChIPseq analyses on SC, RS and ES cell stages. Spearman correlation analyses between the different cell stages showed that there is a strong correlation between replicates of each cell type (0.78, 0.84 and 0.81 respectively for SC, RS and ES). The three datasets strongly correlated with each other (0.74-0.81) The peak calling analysis identified a similar number of peaks and positions for the two replicates of SC (52977 and 53292) and RS (69485 and 74382). Only the peaks common between replicates were considered, we obtained 43856 peaks in SC and 67918 peaks in RS. Due to its advanced state of condensation, the number of peaks was lower for ES and different between replicates (7606 and 1224, with 385 common peaks).

In terms of profile, H3K79me2 shows a strong signal after the TSS (transcription start site) which slowly decreases along the gene body. This profile is similar between the SC, RS and ES stages, and is the typical profile observed in the literature [17]. The signal level decreases between SC and RS and between RS and ES (Fig. 3A). Peak annotation shows that in SC, RS and ES many peaks are annotated as upstream and intragenic regions. In RS, ~13000 peaks are associated with distal intergenic regions (19%) (Fig. 3B). In SC and RS, H3K79me2 enrichment correlates with gene expression level, since highly expressed genes are more enriched in H3K79me2 (Fig. 3C). ES datasets were not further analyzed since the number of common peak was low indicating poor reproducibility.

3.2 Prediction of mouse genome chromatin states in SC and RS cell stages

To understand what are the distal intergenic regions identified in RS, what they are associated with and whether these regions may be distal regulatory regions (enhancers), we decided to compare H3K79me2 with other histone marks using ChromHMM. First, we produced a ChromHMM model using 14 mouse datasets from ENCODE representing 6 different epigenetic marks (Fig. 4A). With this model we obtained 18 different chromatin states. Each chromatin state was characterized by one or more histone marks and was named according to the nomenclature used in ChromHMM publications [16,29,31].

We then focused on the distribution of the different chromatin states at the SC and RS stages (see Fig. 4B). To be able to compare chromatin states between SC and RS, we excluded ZNF/Rpts and Het chromatin stages because they are characterized by H3K9me3 and we did not have this data for SC. Four chromatin states were found proximal to active TSSs states (TSS, TSSFlnk, TSSFlnkD and TSSFlnkU); two states were associated with transcription (Tx and TxWk); four states were enhancer-related (EnhA, EnhG1, EnhG2, and EnhW); two were bivalent states (TssBiv and EnhBiv); two states were repressive (ReprPC and ReprPCWk enriched in H3K27me3) and two states were quiescent (Quies1, Quies2). Similar chromatin states were merged to simplify analysis, EnhG1 and EnhG2 were merged in EnhG, and Quies1 and Quies2 were merged in Quies.

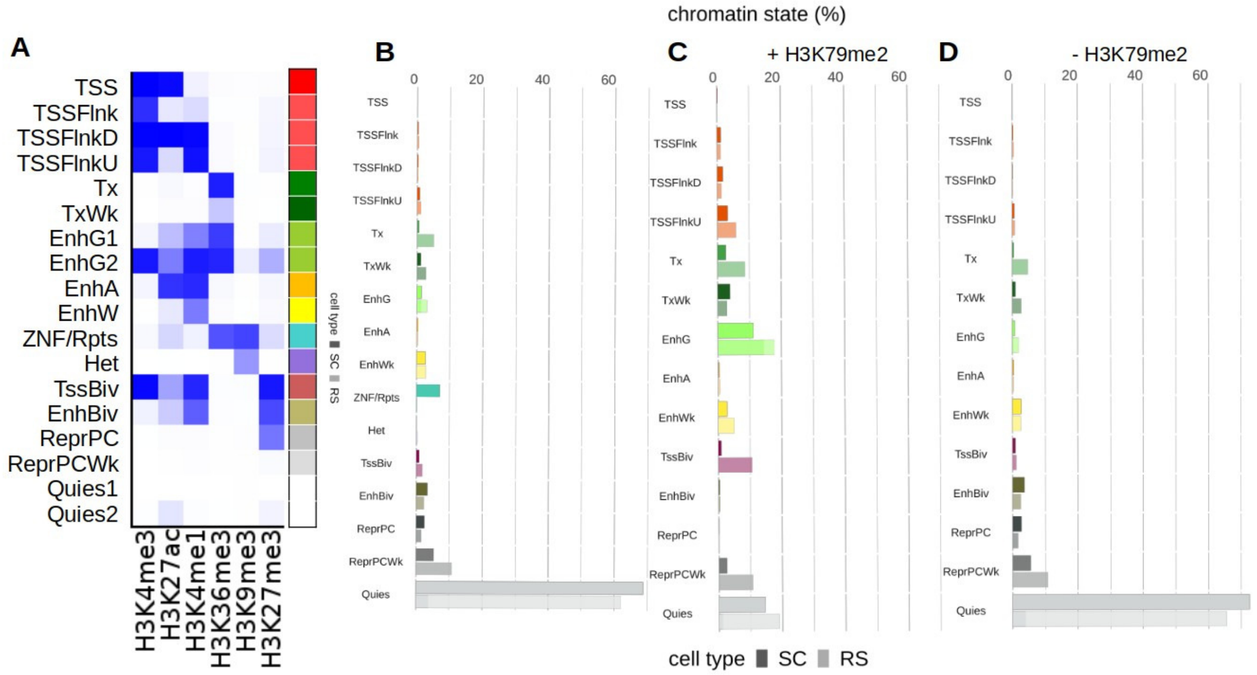


Fig. 4. Prediction of mouse genome chromatin states in SC and RS stages. **A.** ChromHMM model. **B.** Distribution of chromatin states coverage in SC and RS stages. **C**. Distribution of chromatin states with H3K79me2 in SC and RS stages. **D**. Distribution of chromatin states without H3K79me2 in SC and RS stages. TSS : Transcription Start Site, TSSFlnk : Flanking TSS, TSSFlnkD : Flanking TSS downstream, TSSFlnkU : Flanking TSS upstream, TX : Strong transcription, TxWk : Weak transcription, EnhG1 and EnhG2 : Genic enhancers, EnhA : Active enhancer, EnhW : Weak enhancer, ZNF/Rpts : ZNF genes and repeats, Het : Heterochromatin, TSSBiv : Bivalent/Poised TSS, EnhBiv : Bivalent enhancer, ReprPC : Repressed PolyComb, ReprPCWk : Weak repressed polycomb, Quies1 and Quies2 : Quiescent/Low.

To characterize H3K79me2, we integrated H3K79me2 ChIP-Seq datasets with the chromatin states obtained with ChromHMM (Fig. 4C-D). We observed that H3K79me2 was more associated with transcription (6.4% in SC and 11.5% in RS, versus 0.9% and 7.4% without H3K79me2), flanking TSS (6% in SC and 8% in RS versus 1.1% in both SC and RS without H3K79me2), genic enhancer (15% in SC and 20.7% in RS, versus 0.9% and 2.2% respectively in both SC and RS), and with bivalent states, only in RS (2% in SC and 11% in RS versus 4.3% and 3.6%) stages. This confirms that H3K79me2 mark is linked to active transcription [12].

3.3 Impact of H3K79me2 on enhancer-gene expression

We next investigated whether the presence of H3K79me2 at enhancers had an impact on the expression of associated genes. First, we compared the expression level of genes associated with enhancers enriched or not in H3K79me2 according to their distance. In SC and RS, whatever the type of enhancer, the expression level of genes located >100kb from enhancer did not show any specific trend (Fig. 5A). For genic enhancers, the expression level of nearby genes (<100kb from enhancer) was higher than of distal genes. The presence of H3K79me2 at enhancers was associated with a significantly higher level of expression of nearby genes compared to enhancers without the mark except for genic enhancer (wilcoxon test, p-value < 0,05, Fig. 5B). This is consistent with the observations that H3K79me2 is a mark associated with gene expression, both at the gene body and at enhancer regions [32,33].

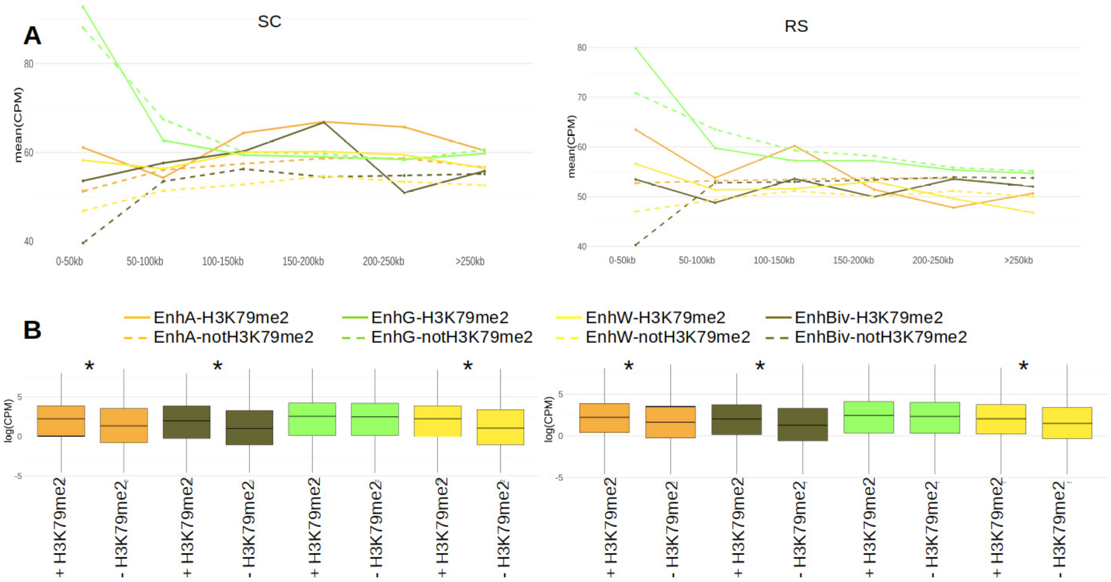


Fig. 5. A. The mean expression level of gene expression (log(CPM)) in function of the distance between gene and enhancer in SC (left) and RS (right). Solid lines correspond to enhancer enriched in H3K79me2 and dashed lines correspond to enhancer not enriched in H3K79me2. B. Expression (log(CPM)) of nearby genes (distance <100kb) in each enhancers category enriched or not in H3K79me2 in SC (left) and RS (right)

3.4 Impact of *Dot1l*-KO on transcription

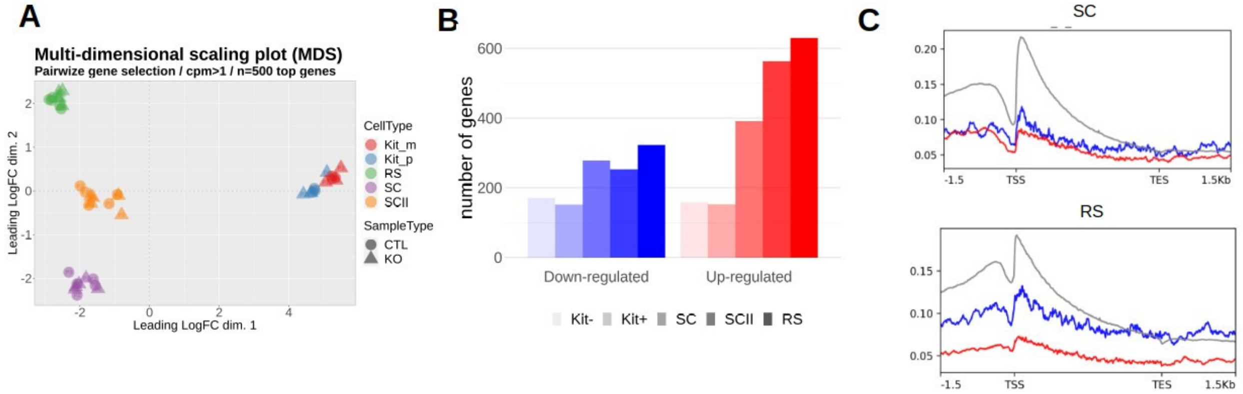


Fig. 6. RNAseq results **A.** Multi-Dimensional Scaling figure between Kit-, Kit+, SC, SCII and RS stages. **B.** Number of deregulated genes identified by differentially expressed genes analysis. **C.** Deregulated gene enrichment in H3K79me2 in SC and RS, upregulated genes (red), with a log(KO/control) > log(1.5) (p-value<5%); downregulated genes (blue), with a log(KO/control) <-log(1.5) (p-value < 5%); not regulated genes (grey)

In order to analyze the impact of *Dot1l*-KO on transcription, we performed RNAseq analysis using data from 5 stages of spermatogenesis: undifferentiated spermatogonia (Kit-), differentiated spermatogonia (Kit+), primary spermatocytes (SC), secondary spermatocytes (SCII) and round spermatids (RS). The multi-dimensional scaling (MDS) representation of gene expression shows that the different spermatogenesis cell stages are distinct (Fig. 6A). Spermatogonia have a different transcriptional dynamic in comparison with spermatocytes and spermatids explained by the first dimension. We performed differential expression analysis between *Dot1l*-KO and control samples and observed that in spermatogonia (Kit- and Kit+), the number of deregulated genes is substantially lower than in the other stages. From SC stage, the number of upregulated genes (159, 153, 392, 563, 630 in Kit-, Kit+, SC, SCII and RS respectively) is higher than downregulated ones (171, 152, 278, 253, 323 in Kit-, Kit+, SC, SCII and RS respectively)(Fig. 6B). This was recently described in other cell types such as B and T cells [34,35]. This observation is unexpected because the presence of H3K79me2 positively correlates with the active transcription of thousands genes. We therefore expected to have more than 5% deregulated genes, and more downregulated than upregulated genes when *Dot1l* is knocked-out. We investigated H3K79me2 enrichment at deregulated genes and found that gene deregulation did not strictly depend on H3K79me2 (Fig. 6C).

3.5 Characterization of chromatin environment of deregulated genes

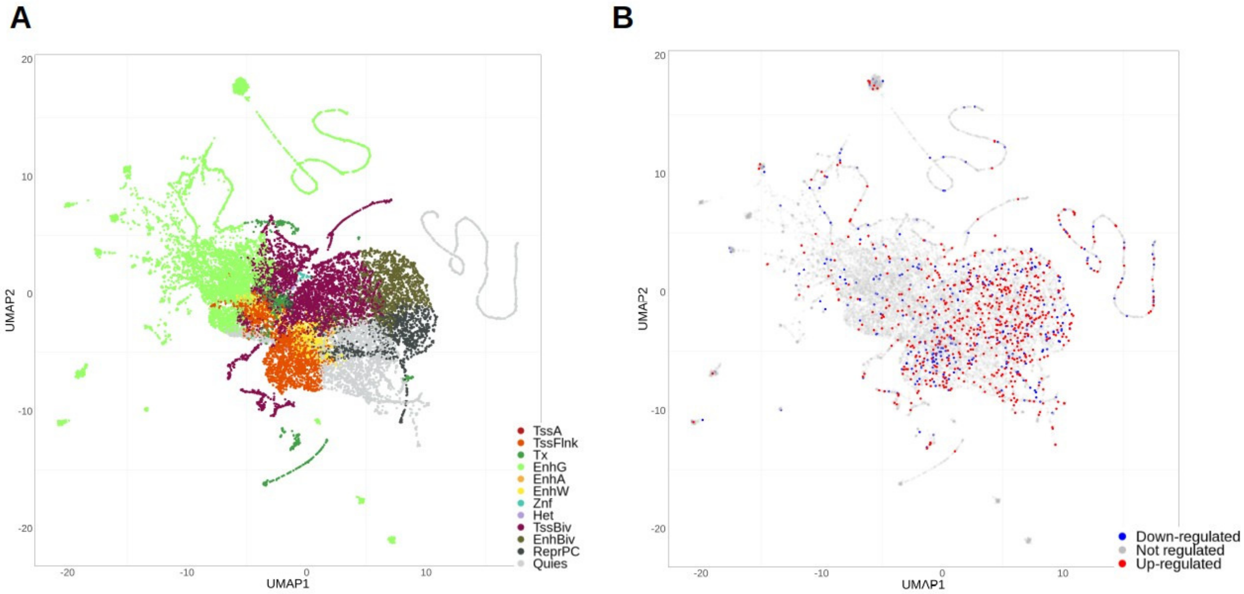


Fig. 7. Dimensionality reduction (UMAP) of the predominant chromatin state of genes for RS. Genes are represented as dots. **A.** Color-code based on the predominant chromatin state present in the gene environment **B.** Color-code based on the gene expression level defined in RNAseq analysis.

We were therefore interested in knowing if a specific chromatin environment could be responsible for the observed deregulation. To do this, we used the chromatin states defined by ChromHMM to look at the chromatin environment of the 20,639 genes included in our study. We assigned to each gene a predominant chromatin state based on the chromatin state showing the highest coverage within the TSS±3kb region (Fig. 7A). We correlated the predominant chromatin state to the genes found significantly deregulated in *Dot1l*-KO SC and RS (Fig. 7B). In SC, 21.6% of upregulated genes and 31% of downregulated genes are associated to bivalent states (versus 16.7% for total genes), 45% and 27% of upregulated and downregulated genes with quiescent state (versus 24% for total), and 11% and 8% with repressor state (versus 5%). In RS, 36% of upregulated genes and 24% of downregulated genes are associated to bivalent states (29% for total), 26% and 19% of upregulated and downregulated genes with a quiescent state (versus 15%) and 13% and 10% with repressed state (versus 6%). Overall, in both SC and RS, we found that predominant chromatin states in the upregulated genes are characterized by H3K27me3, a repressive mark. We could therefore hypothesize that DOT1L could have a role related to H3K27me3 as recently observed by Kwesi-Maliepaarda paper [34]. In their model, DOT1L downregulates *EZH2*, a gene coding for a protein of PRC2 (polycomb repressor 2), responsible for H3K27 trimethylation, and associated with gene repression. In *Dot1l*-KO germ cells, *EZH2* is not downregulated ; more work will therefore be needed to identify the link between DOT1L and H3K27me3-mediated repression.

4 Conclusion

In SC and RS male germ cells, as in other cell types, H3K79me2 is a broad mark associated with gene expression. We built a ChromHMM model using 6 histone marks and identified 18 different chromatin states with which H3K79me2 ChIP-Seq was integrated. We found that H3K79me2 is mainly associated to promoter, genic enhancer, weak enhancer and bivalent states in SC and RS. Focusing on enhancers states, we observed that the presence of H3K79me2 at enhancers is associated with a high level of expression of nearby genes.

The study of the impact of *Dot1l*-KO on gene expression shows ~800 upregulated genes and ~250 downregulated genes, and no strict correlation between deregulation and the presence of H3K79me2 mark. The analysis of the chromatin environment at deregulated genes shows that upregulated genes are predominantly associated with bivalent states and repressed states in SC and RS. These states are characterized by the presence of a repressive histone mark, H3K27me3, suggesting a link between DOT1L and H3K27me3, independent of H3K79 methylation.

References

1. Soumillon, M. *et al.* Cellular source and mechanisms of high transcriptome complexity in the mammalian testis. *Cell Rep.* **3**, 2179–2190 (2013).
2. Green, C. D. *et al.* A Comprehensive Roadmap of Murine Spermatogenesis Defined by Single-Cell RNA-Seq. *Dev. Cell* **46**, 651-667.e10 (2018).
3. Maezawa, S., Yukawa, M., Alavattam, K. G., Barski, A. & Namekawa, S. H. Dynamic reorganization of open chromatin underlies diverse transcriptomes during spermatogenesis. *Nucleic Acids Res.* **46**, 593–608 (2018).
4. Blanco, M. & Cocquet, J. Genetic Factors Affecting Sperm Chromatin Structure. *Adv. Exp. Med. Biol.* **1166**, 1–28 (2019).
5. Rathke, C., Baarends, W. M., Awe, S. & Renkawitz-Pohl, R. Chromatin dynamics during spermiogenesis. *Biochim. Biophys. Acta BBA - Gene Regul. Mech.* **1839**, 155–168 (2014).
6. Sonnack, V., Failing, K., Bergmann, M. & Steger, K. Expression of hyperacetylated histone H4 during normal and impaired human spermatogenesis. *Andrologia* **34**, 384–390 (2002).
7. Fenic, I., Sonnack, V., Failing, K., Bergmann, M. & Steger, K. In vivo effects of histone-deacetylase inhibitor trichostatin-A on murine spermatogenesis. *J. Androl.* **25**, 811–818 (2004).
8. Shirakata, Y., Hiradate, Y., Inoue, H., Sato, E. & Tanemura, K. Histone h4 modification during mouse spermatogenesis. *J. Reprod. Dev.* **60**, 383–387 (2014).
9. Schon, S. B. *et al.* Histone modification signatures in human sperm distinguish clinical abnormalities. *J. Assist. Reprod. Genet.* **36**, 267–275 (2019).
10. Dottermusch-Heidel, C. *et al.* H3K79 methylation directly precedes the histone-to-protamine transition in mammalian spermatids and is sensitive to bacterial infections. *Andrology* **2**, 655–665 (2014).
11. Kim, W., Choi, M. & Kim, J.-E. The histone methyltransferase Dot1/DOT1L as a critical regulator of the cell cycle. *Cell Cycle* **13**, 726–738 (2014).
12. Barski, A. *et al.* High-Resolution Profiling of Histone Methylations in the Human Genome. *Cell* **129**, 823–837 (2007).
13. Jones, B. *et al.* The Histone H3K79 Methyltransferase Dot1L Is Essential for Mammalian Development and Heterochromatin Structure. *PLoS Genet.* **4**, (2008).
14. Moretti, C. *et al.* SLY regulates genes involved in chromatin remodeling and interacts with TBL1XR1 during sperm differentiation. *Cell Death Differ.* **24**, 1029–1044 (2017).
15. Ernst, J. & Kellis, M. ChromHMM: automating chromatin-state discovery and characterization. *Nat. Methods* **9**, 215–216 (2012).
16. Ernst, J. & Kellis, M. Chromatin-state discovery and genome annotation with ChromHMM. *Nat. Protoc.* **12**, 2478–2492 (2017).
17. Vlaming, H. & van Leeuwen, F. The upstreams and downstreams of H3K79 methylation by DOT1L. *Chromosoma* **125**, 593–605 (2016)
18. Köster, J. & Rahmann, S. Snakemake—a scalable bioinformatics workflow engine. *Bioinformatics* **28**, 2520–2522 (2012).
19. Bushnell, B. BBMap: A Fast, Accurate, Splice-Aware Aligner. in (2014).
20. Dobin, A. *et al.* STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29**, 15–21 (2013).
21. Love, M. I., Huber, W. & Anders, S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol.* **15**, 550 (2014).
22. Robinson, M. D., McCarthy, D. J. & Smyth, G. K. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinforma. Oxf. Engl.* **26**, 139–140 (2010).
23. Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. *Nat. Methods* **9**, 357–359 (2012).
24. Zhang, Y. *et al.* Model-based Analysis of ChIP-Seq (MACS). *Genome Biol.* **9**, R137 (2008).
25. Yu, G., Wang, L.-G. & He, Q.-Y. ChIPseeker: an R/Bioconductor package for ChIP peak annotation, comparison and visualization. *Bioinforma. Oxf. Engl.* **31**, 2382–2383 (2015).
26. Bryant, J. M. *et al.* Characterization of BRD4 during mammalian postmeiotic sperm development. *Mol. Cell. Biol.* **35**, 1433–1448 (2015).
27. Hammoud, S. S. *et al.* Chromatin and Transcription Transitions of Mammalian Adult Germline Stem Cells and Spermatogenesis. *Cell Stem Cell* **15**, 239–253 (2014).
28. Zuo, X. *et al.* The histone methyltransferase SETD2 is required for expression of acrosin-binding protein 1 and protamines and essential for spermiogenesis in mice. *J. Biol. Chem.* **293**, 9188–9197 (2018).
29. ENCODE Project Consortium. An integrated encyclopedia of DNA elements in the human genome. *Nature* **489**, 57–74 (2012).
30. Konopka, Tomasz. *umap: Uniform Manifold Approximation and Projection* (2022)
31. van der Velde, A. *et al.* Annotation of chromatin states in 66 complete mouse epigenomes during development. *Commun. Biol.* **4**, 1–15 (2021).
32. Godfrey, L. *et al.* DOT1L inhibition reveals a distinct subset of enhancers dependent on H3K79 methylation. *Nat. Commun.* **10**, 1–15 (2019).
33. Ferrari, F. *et al.* DOT1L-mediated murine neuronal differentiation associates with H3K79me2 accumulation and preserves SOX2-enhancer accessibility. *Nat. Commun.* **11**, 5200 (2020).
34. Kwesi-Maliepaard, E. M. *et al.* The histone methyltransferase DOT1L prevents antigen-independent differentiation and safeguards epigenetic identity of CD8+ T cells. *Proc. Natl. Acad. Sci. U. S. A.* **117**, 20706–20716 (2020).
35. Aslam, M. A. *et al.* Histone methyltransferase DOT1L controls state-specific identity during B cell differentiation. *EMBO Rep.* **22**, e51184 (2021).

Insertion of *Alu* elements impacts sequence-mediated nucleosome positioning

Jérémy Barbier[1,2], Fabien Sassolas[1,2], Cédric Vaillant[1], Jean-Nicolas Volff[2], Frédéric G. Brunet[2] and Benjamin Audit[1]

[1] ENS de Lyon, CNRS, Laboratoire de Physique, F-69342 Lyon, France

[2] Institut de Génomique Fonctionnelle de Lyon, Ecole Normale Supérieure de Lyon, CNRS UMR 5242, Univ Claude Bernard Lyon 1, F-69364 Lyon, France

Corresponding author: benjamin.audit@ens-lyon.fr

Abstract *A physical model for the formation of nucleosomes revealed that the DNA sequence codes for regions inhibiting the formation of nucleosomes. More than* 1.6 *million nucleosome inhibiting energy barriers (NIEBs) were delineated in the human genome and corroborated as* in vivo *nucleosome depleted regions (NDRs) in experimental nucleosome occupancy profiles. NIEBs are bordered on each side by compact arrays of 2-3 nucleosomes; one third of the human is covered by this intrinsic nucleosomal organization. Mutation profiles and GC content at NIEB loci presented modulations phased with the compact nucleosomal arrays and the comparison between interspecies and intraspecies divergence rates suggested that nucleosome positioning around these intrinsic NDRs was under selection. In all vertebrates analyzed, similar sequence properties were observed indicating that NIEBs with compacted neighboring nucleosomes are in fact ubiquitous. These results suggest the existence of evolutionary mechanisms linked to the intrinsic nucleosomal organization.*

We propose an evolutionary scenario in human and chimpanzee where the insertion of Alu transposable elements give rise to new NIEBs based on the observation that (i) a majority of Alu elements are inserted at NIEB borders, (ii) the most recently inserted Alu elements are the closest to the NIEBs, (iii) human insertion sites of polymorphic Alu elements show no preferential positioning regarding NIEBs and (iv) the species-specific Alu elements are mainly positioned at the borders of species-specific NIEBs. We identify transposable elements (TEs) in mouse and pig that are good candidates to follow a similar model of NIEB creation by insertion of TEs. The dynamics of transposition would thus be a general mechanism of chromatin evolution that could explain the evolutionary success of some TE families. This work allows to better apprehend the evolutionary mechanisms responsible for the genome-wide intrinsic nucleosomal organization.

Keywords DNA sequence-encoded nucleosome ordering; nucleosome depleted regions; Alu elements; chromatin evolution.

1 Introduction

In most Eukaryotes, short chromosome segments (∼ 150 DNA base pairs) wrap around cores of 8 histone proteins to form nucleosomes as the elementary structural unit of DNA folding within chromatin [1]. The control of nucleosome positioning and occupancy is critical for the regulation of gene transcription, DNA replication, repair and recombination [1, 2]. We developed a simple physical model of nucleosome occupancy based on the computation of the free energy cost of bending a DNA fragment of a given sequence from its natural curvature to the final superhelical structure around the histone core from which, given an average nucleosome coverage (chemical potential), the nucleosome occupancy profile can be derived [3]. Combining the nucleosome occupancy probability profile and the original energy profile, we identify nucleosome inhibiting energy barriers (NIEBs) as the genomic loci where the energy is high enough to induce a nucleosome-depleted region in the nucleosome occupancy profile [3–6]. This methodology was developed based on yeast experimental nucleosome positioning data. When applied to human, the method delineated an impressive ∼ 1.6 million NIEBs, demonstrating that NIEBs are an important feature of the human genome [7]. Importantly, we also observe that the model predictions around NIEBs are in very good agreement with *in vitro* nucleosome occupancy data [8]. Not only is a very low nucleosome occupancy observed within the NIEBs but the compact positioning of 2-3 nucleosomes (nucleosome repeat length (NRL) $<$ 160 bp) at each NIEB border predicted by the physical model is also observed in the experimental data. The physical model indeed captures intrinsic sequence-dependent nucleosome positioning signal as we also observed for *in vitro* data for the yeast genome [3, 5, 6]. Average *in vivo* nucleosome occupancy profiles [8, 9] are also extremely consistent with the model predictions at NIEB loci, showing that intrinsic NIEBs and the 4-6 well positioned nucleosomes at their borders are also meaningful to describe *in vivo* chromatin organization. Hence, the 1kb-sized regions around NIEBs (one NIEB and its 4-6 flanking nucleosomes) covering 38 % of the human genome, correspond to regions of intrinsic nucleosome positioning that are not subject to chromatin remodeling [7]. These regions are thus of particular interest to study the nucleosome-associated genome evolution such as mutational patterns and transposable element (TE) insertion. We observed that these intrinsic NDRs, together with the flanking nucleosomes, are actually encoded in the GC content and are widely distributed along the 22 human autosomes in GC-rich and GC-poor isochores, in early and late replicating regions, in intergenic and genic regions [7]. By comparing rates of inter- and intraspecies divergences, we have brought evidence of complex patterns of positive and negative selection that depend on the local GC content to maintain some optimal difference in GC composition between the AT-rich intrinsic NDRs and the GC-rich well-positioned first flanking nucleosomes. Intrinsic chromatin organization encoded by NIEBs is ubiquitous across vertebrates as (i) a significant density of NIEBs $\sim 0.6-0.7$ kb^{-1} was observed in all vertebrate genomes analyzed (compared to a density < 0.01 kb^{-1} for a random sequence with equal proportions of A, G, C and T) and (ii) oscillation of the GC profiles nearby NIEBs in these genomes is remarkably phased with the one in human [10].

Transposable elements constitute a ubiquitous component of vertebrate genomes; they constitute from $\sim$ 6% (Tetraodon, a pufferfish with compact genome) to more than half of the genome in zebrafish and opossum, for example [11]. Initially, TEs were described as selfish DNA that simply take advantage of the cell machinery of their host to multiply and populate the genome as innocent bystanders, mostly without any phenotypic effect [12]. TEs are now recognized as major drivers of gene and genome evolution, with roles in biological diversity and speciation [13–15]. Despite the central role that TEs play in shaping the evolution of eukaryotes, the coupling between the nucleosomal array and the dynamics of TEs has not been fully evaluated. In human, it was indeed proposed that there were no preferential nucleosome positioning over most of the genome [16], preventing the analysis of TE integration sites in relation to nucleosomal positioning. However, using *in vitro* reconstruction of chromatin template with well positioned nucleosomes, it was shown that there exists a dependence on chromatin structure for retroviral integration site selectivity [17, 18]. The observation that in metazoans, the 3D genome folding as measured using chromatin conformation capture experiments correlates with the association of repetitive elements of the same family such as *Alu* in human [19], further highlights the importance to fully characterise the relationship between TEs and chromatin organization. Interestingly, we found that many (52%) *Alu* elements (high copy number non-autonomous TEs found only in primates) were inserted flanking a NIEB [7, 10]. Note that since the nucleosome occupancy model used for NIEB detection was developed on yeast data, the fitting of the model parameters was not influenced by the specific base composition of *Alu* elements. The orientation of the *Alu* elements was strongly dependent on which NIEB side they were inserted, excluding a purely random process. *Alu* sequences are mainly sense at the NIEB 5' end and antisense at the NIEB 3' end, so that the body of the *Alu* element is external to the NIEB. This remarkable positioning results from the matching of the polyA tail of the sense (resp. antisense) *Alu* with the polyA (resp. polyT) located at the edges of some of the predicted NIEBs, leading to an asymmetric distribution of polyA and polyT at *Alu*-associated NIEB borders [7, 10]. Since the terminal polyA at the 3' end of *Alu* plays a critical role in the amplification mechanism [20, 21], this raised the question of the link between NIEBs and *Alu* insertions. The fact that a majority (61%) of NIEBs are free on either side of detectable *Alu* elements initially suggested that NIEBs would pre-exist *Alu* integration [7]. However, *Alu* might be involved in the formation of some but not all NIEBs, or other TEs might be associated to NIEBs, suggesting a convergent mechanism of NIEBs formation by TE insertion. Here we compare the history of *Alu* element insertions along the primate evolutionary tree to NIEB predictions in human and chimpanzee, in order to assess to which extent NIEBs simply constitute a favorable substrate for *Alu* insertions or if these insertions contribute

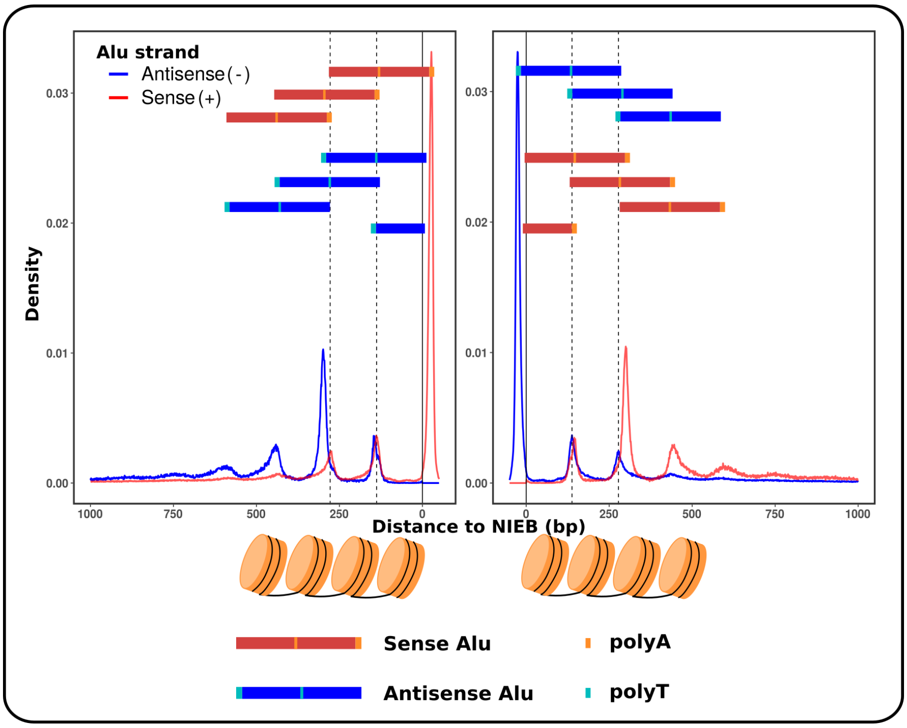


Fig. 1. Distance of the terminal polyA of the 1 078 322 Alu elements on human autosomes to the closest border of nucleosomal barriers (abscissa 0 bp). Alu were classified into 4 categories depending on their orientation (sense Alu in red, antisense Alu in blue) and to which NIEB side they are closest (closest to a NIEB 5' border on the left panel or closest to a NIEB 3' border on the right panel). Each curve is normalized with the total number of Alu of corresponding category (395 076 sense Alu and 142 504 antisense Alu upstream of a NIEB (left panel); 141 548 sense Alu and 398 976 antisense Alu downstream of a NIEB (right panel); 218 Alu were not considered as they fall in a discarded inter-NIEB region containing not sequenced nucleotides (Ns)). Vertical dashed lines mark distances 138 bp and 278 bp from the NIEB borders. The compact arrays of nucleosomes positioned at NIEB borders are drawn under the graph.

to the formation and/or maintenance of NIEBs. In other word, we will address if there is anything else but success of selfishness of the *Alu* amplification to explain the astonishing *Alu* expansion in primates.

2 Results

2.1 Alu elements have specific positioning at NIEBs borders

We first further characterize the positioning of Alu elements relative to NIEBs borders. Alu are dimeric elements, with two GC-rich arms separated by a short polyA sequence and with a longer polyA sequence at their 3' end. To position an Alu element regarding NIEBs, we used the middle of its terminal polyA as a reference position. Figure 1 represents the distributions of these reference positions relative to NIEB borders. Upstream of NIEBs, we observe that sense Alu elements have a strong preferential positioning at NIEBs borders, with most of them being placed with their terminal polyA co-localizing with the NIEB 5' border, as previously described [7, 10]. Interestingly, in this configuration the internal polyA of the elements is placed at the first linker DNA position. Secondary preferential positions are also observed at distance 138 bp and 278 bp from the NIEBs borders, with both the terminal and internal polyAs co-localizing with inter-nucleosomal sequences. The preferential Alu positioning with the polyA in between nucleosomes is also observed for antisense elements, although they represent only a quarter of total Alu elements upstream of NIEBs. Finally, there is a symmetry between the distribution of sense Alu upstream of NIEBs and antisense Alu downstream of NIEBs, and vice-versa.

Alu preferential positioning at NIEBs border could come from several mechanisms:

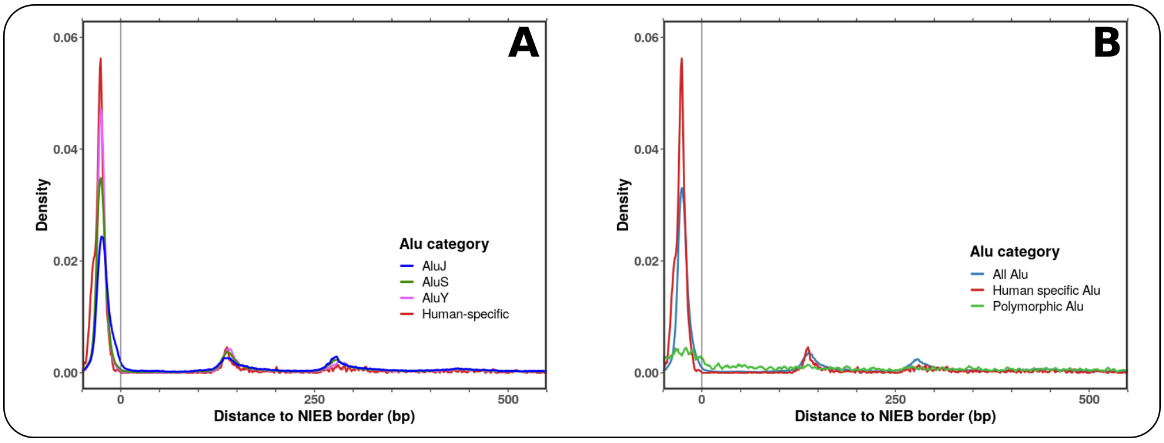


Fig. 2. Distances to the closest nucleosomal barrier border of Alu terminal polyA for different Alu families and of the Alu insertion sites delineated as the location of polymorphic Alu elements from the 1000 Genomes Project. (A) Distance distribution for 95 859 AluJ, 234 121 AluS, 49 188 AluY and 3160 human-specific Alu are represented by the blue, green, pink and red curves, respectively. The analysis is restricted to antisense Alu, less than 2000 bp downstream of a NIEB for which family classification was possible. Alu elements identified as human-specific are included in other categories (mainly in the AluY category). (B) Distance distributions are computed for antisense Alu elements positioned at less than 2000 bp upstream of a NIEB: all 392 807 such Alu in blue, 3160 human-specific Alu in red and 3610 Alu insertion sites of human polymorphic Alu elements in green. Note that x-axis only ranges from -50 bp to 550 bp, because profiles obtained between 550 and 2000 are flat and nearly equal to 0.

— A strong counter-selection of ”misplaced” insertions because they would disrupt the chromatin structure.

— Preference for AT-rich, inter-nucleosomal sequences for the insertion of Alu elements because of the insertion mechanism.

— Creation of NIEBs in phase with newly inserted Alu elements, thanks to the inhibitory effect of polyA sequences on nucleosome positioning.

2.2 The latest Alu insertions show strongest constraints on positioning

To explore the hypothesis of a counter-selection of Alu insertions disrupting the chromatin structure, we classified the Alu elements according to their ages (inserting time in the genome). Indeed, we expect the constraints on Alu positioning to be stronger on the oldest categories, simply because the time during which selection has occurred is longer. We used the classification of Alu elements in three main families, with AluJ (26 %) being the oldest elements in the human genome, followed by AluS (59 %), and then AluY (12 %) being the youngest family. We also considered the set of Alu that have been identified as human specific (0.7 %) by Tang et al. [22, 23]. This last category represent the youngest Alu elements of the human genome, as they were inserted after human-chimpanzee divergence. Note that the elements in this category mainly belongs to the AluY family. However, they represent only a small fraction of it (about 6%).

We observe that constraints on Alu positioning are stronger on younger insertions (Figure 2 - A). Indeed, there is a gradient in the proportion of elements positioned at the internal border of NIEBs, with the highest proportion for human-specific Alu, followed by AluY, then AluS and finally AluJ. Thus, the younger an Alu element is, the more its positioning is constrained at NIEBs borders, which is not compatible with the counter-selection hypothesis developed above. Therefore, ”misplaced” insertions do not seem to be purified by selection. The next hypothesis is that the insertion mechanism of Alu elements puts them directly where we find them now.

2.3 Human polymorphic Alu insertion sites are not biased toward NIEBs borders

Alu elements are inserted through the binding of their terminal polyA to a pre-existing A-rich sequence in the genome [21]. Yet, an increase in the polyA distribution has been shown at the internal border of NIEBs and at first linker position [7, 10]. This increase could make these regions preferential insertion platforms for Alu elements, especially as they are also inter-nucleosomal, potentially more accessible to the transposition machinery. To explore this hypothesis, we established the distribution of Alu insertion sites delineated as the location of polymorphic Alu elements extracted from the 1000 Genome Project data [24, 25]. The Alu are present in a human sub-population but are absent from the reference genome. Following our assumption, the distribution of insertion sites should be similar to the one observed for the Alu elements present in the human reference genome.

On Figure 2 - B, we can see that insertion sites of the polymorphic Alu element (in green) show only a very little preference for NIEBs, with only 10.8% of them sitting at the internal border of a NIEB, as compared to the 40% of Alu positioned with their polyA at this locus. This result invalidates the assumption that the insertion mechanism of Alu elements is responsible for their distribution at NIEBs borders.

2.4 Alu insertion creates new nucleosomal barriers

The results presented in sections 2.2 and 2.3 not only invalidated two of our three hypothesis, they also tend to confirm the third one, which stipulates that Alu element insertions would be at the origin of new NIEBs. Indeed, polymorphic elements are not present in human reference genome, and were not taken into account in the human NIEBs prediction. Thus, it seems that NIEBs are found at Alu sites only when the Alu are present in the analyzed sequence. This result is in favor of NIEBs created by the Alu insertions. Moreover, in this model, NIEBs would be formed on the terminal polyA of Alu elements. It is known that this polyA is susceptible to shrinkage when the Alu gets older [26]. Yet, oldest Alu are also less positioned at NIEBs borders. An Alu insertion could then create a NIEB at the terminal polyA of the element. After the insertion, the shrinkage of the polyA could progressively weaken the corresponding NIEB border, which would explain the relaxed constraints on older Alu element positioning at NIEBs borders. To confirm this model of NIEB creation, we compared nucleosome occupancy at Alu sites with and without the inserted Alu sequence using a comparative genomics approach described in section 4.2.

We see that the nucleosome occupancy predicted by our model corroborates our hypothesis of NIEB creation by the insertion of Alu (Figure 3). Indeed, the predicted occupancies in the presence of Alu (in blue) are very different from the ones obtained without the Alu elements (in red). The polyA of the elements (around position 0 bp) systematically form a zone where the occupancy is very low, bordered by two to three well positioned nucleosomes. This nucleosomal organization is characteristic of NIEBs. We also note that the predicted occupancies on Alu elements are particularly high, reflecting the compatibility between these elements and nucleosome positioning. Finally, we observe that the occupancy pattern predicted from our physical model of nucleosome formation is consistent whether or not a NIEB was detected at the insertion site and/or at the border of the Alu element. It is thus clear that the absence of detected NIEB for the Alu of the C and D panels on Figure 3 comes from false negative in NIEBs detection. Hence, a vast majority of human-specific Alu are at the origin of human-specific NIEBs.

3 Discussion

The results presented here demonstrate that the insertion of Alu elements have a strong effect on nucleosomal positioning. Indeed, these elements are at the origin of new NIEBs at the border of which nucleosome are well positioned. This association between Alu elements and NIEBs could have had a role in the spreading of these elements among primates genomes, as they could have been used as spreading factors for NIEBs. Our results also highlight the importance of the terminal polyA of Alu element in NIEB formation. Yet, a lot of SINEs are terminated by polyA, leading to the question of a general interaction between transposable elements and NIEBs. For example, in mouse, B1 elements derives from the same 7SLRNA as Alu elements in primates. The main difference between B1 and Alu is that B1 is a monomer whereas Alu is a dimer. In that way, B1 could be seen as a "half Alu". We found that at the border of NIEBs, B1 are positioned with the same constraints as the ones detailed here for Alu (Figure 1), namely a preference for a positioning with the terminal polyA at the internal border of NIEBs, or at least in an inter-nucleosomal sequence. Thus, B1 elements in mouse could also be involved in NIEBs formation in this genome. In pig, Pre0_SS elements are GC-rich

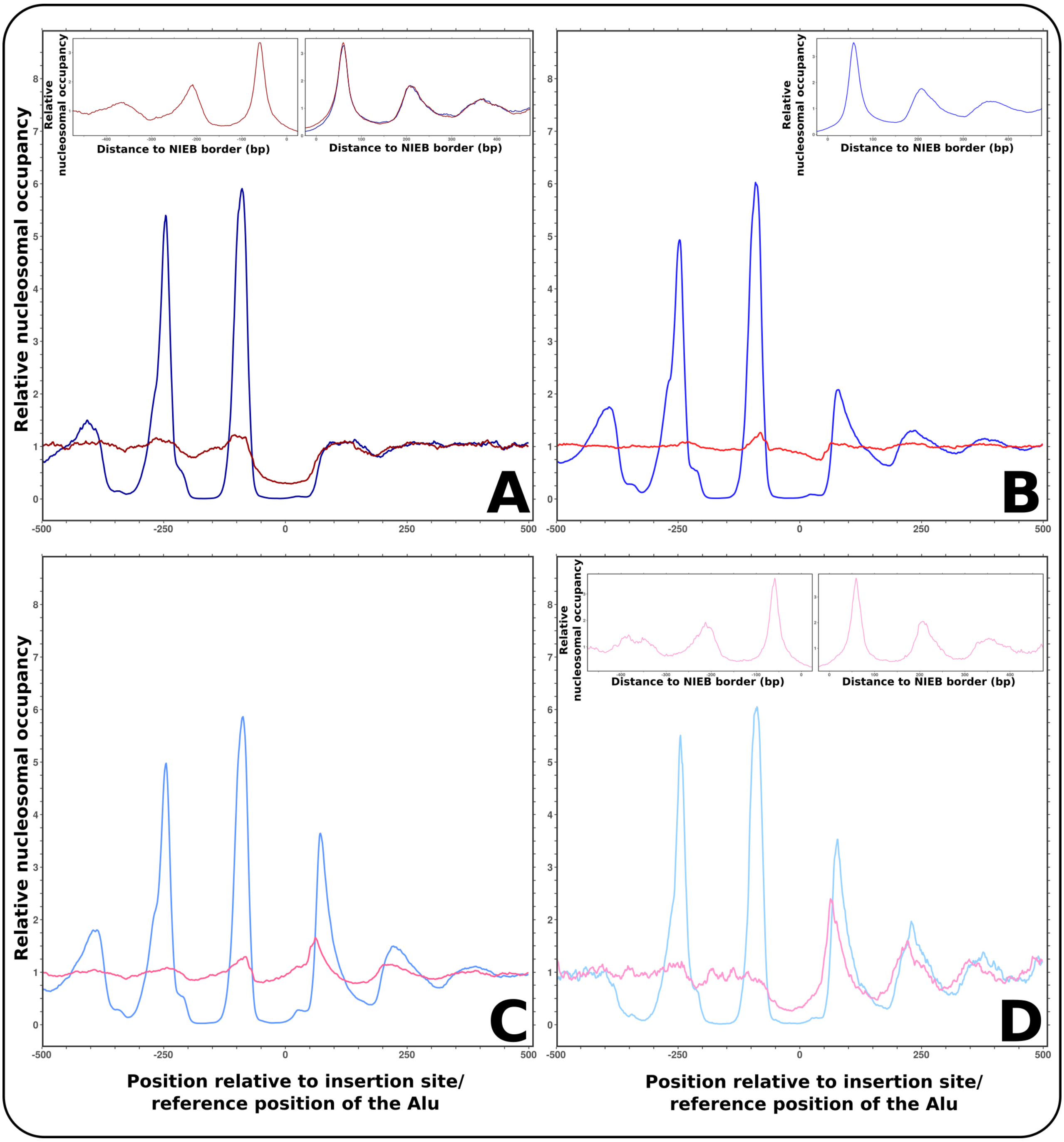


Fig. 3. Mean predicted nucleosomal occupancy at human-specific Alu sites in human and chimpanzee. Red curves represent the profiles obtained in chimpanzee (without Alu element), blue curves the profiles obtained in human (with Alu element). A, B, C and D figures represent respectively cases 1, 2, 3 and 4 described in section 4.2. On each graph, the 0 of the x-axis correspond to reference position for the Alu element in human, and the corresponding insertion site positioned in chimpanzee retrieved through the alignment of the 2x100 bp flanking the human Alu in the chimpanzee genome. Boxes in A, B and D figures represent the same profiles but using as reference position the borders of NIEBs. All these profiles have been normalized with the genomic mean of the signal.

SINEs comprising a terminal polyA. Their size (about 250 bp) and base composition is close to the ones of Alu elements, besides the internal polyA separating the two GC-rich arms of the Alu. Here again, we found that the positioning of these elements at NIEBs borders reproduce the one observed for Alu (and B1) elements. These elements are specific to pigs, and they have a totally different origin as the Alu (they derive from a tRNA-Glu sequence). Thus, it seems that several TE families in several species show similar effect on nucleosomal positioning. This leads to the possibility of a general interaction between NIEBs and TEs, with implication in the evolutionary success of certain TE families. Interestingly, it was observed that regions of preferential DNA replication initiation common to all cell lines were enriched in NIEBs compared to cell line specific replication initiation regions [27, 28], suggesting that TE insertion could contribute to the specification of these constitutive replication initiation regions.

4 Methods

4.1 Reference genomes and annotations

Human (hg38 and hg19) and chimp (panTro5) genomes were dowloaded on UCSC database (http://hgdownload.soe.ucsc.edu/downloads.html). The annotation of transposable elements used for these two species was the one obtained with RepeatMasker software and provided by the UCSC database. For NIEB prediction, we used the physical model based on sequence-dependent DNA bending properties described in [3–5] to detect 1 745 801, 1 718 916 and 1 733 364 NIEBs respectively on human (hg38 and hg19) and chimp autosomes. *In silico* nucleosome density were obtained with the same method as described in [7] to model high nucleosome coverage *in vivo*. Insertion sites of polymorphic Alu elements were extracted from the set of human polymorphisms detected regarding hg19 genome version in the phase 3 of the 1000 Genomes Project [24, 25]. They were compared to the NIEB prediction obtained for this version of the human genome. NIEBs positions are available at https://perso.ens-lyon.fr/benjamin.audit/Vertebrate_NIEBs/. Predicted nucleosome occupancy profiles can be obtained from the corresponding author.

4.2 Human-specific Alu elements and corresponding insertion sites in chimp

To retrieve ancestral insertion site of human-specific Alu elements, we aligned the flanking regions of each human-specific Alu elements on the chimp genome. 100 bp on each side of each Alu elements were taken and put in fastq format to mimic a paired-end sequencing dataset. These ”read pairs” were then aligned on chimp genome using bwa-mem (http://bio-bwa.sourceforge.net/bwa.shtml). Only alignments that were properly paired, with the two reads slightly overlapping or being side by side were kept for further analysis. Also, every alignment comprising an indel were discarded, to ensure that the ±100 bp around the ancestral insertion site were identical as the ±100 bp around the human Alu element (except for potential SNVs). Human-specific Alu associated with an ancestral insertion site were then separated in 4 groups according to the presence/absence of NIEB at the reference position of Alu element in human and at corresponding insertion site in chimp :

1. A NIEB is present both with (in human) and without (in chimp) Alu element (572 cases, 12.1%)
2. A NIEB is present only associated to the Alu element in human (2229 cases, 47.1%)
3. There is no NIEB neither with nor without the Alu element (1831 cases, 38.7%)
4. A NIEB is present only without the Alu element in chimp (96 cases, 2.0%)

The profiles presented in Figure 3 were obtained by taking the mean occupancy profile of each group in human and chimp, aligned on the reference position.

Acknowledgements

This work was funded by the Agence Nationale de la Recherche (ANR-20-CE12-013), and the ENS de Lyon “Projets Emergents” program. J.B. acknowledges support from the PhD funding program of ENS de Lyon. The authors thank the French CNRS network GDR ”Architecture et Dynamique du Noyau et des Génomes” (ADN&G) for stimulating workshops.

References

[1] Zlatanova, J. & Van Holde, K. E. (2016). *Molecular biology : structure and dynamics of genomes and proteomes*. Garland Science, Taylor and Francis Group, New-York.

[2] Arneodo, A., Vaillant, C., Audit, B., Argoul, F., d'Aubenton-Carafa, Y. & Thermes, C. (2011). Multi-scale coding of genomic information: From DNA sequence to genome structure and function. *Phys. Rep.* 498, 45–188.

[3] Chevereau, G., Arneodo, A. & Vaillant, C. (2011). Influence of the genomic sequence on the primary structure of chromatin. *Front. Life Sci.* 5, 29–68.

[4] Vaillant, C., Audit, B. & Arneodo, A. (2007). Experiments confirm the influence of genome long-range correlations on nucleosome positioning. *Phys. Rev. Lett.* 99, 218103.

[5] Milani, P., Chevereau, G., Vaillant, C., Audit, B., Haftek-Terreau, Z., Marilley, M., Bouvet, P., Argoul, F. & Arneodo, A. (2009). Nucleosome positioning by genomic excluding-energy barriers. *Proc Natl Acad Sci U S A* 106, 22257–22262.

[6] Vaillant, C., Palmeira, L., Chevereau, G., Audit, B., d'Aubenton Carafa, Y., Thermes, C. & Arneodo, A. (2010). A novel strategy of transcription regulation by intragenic nucleosome ordering. *Genome Res* 20, 59–67.

[7] Drillon, G., Audit, B., Argoul, F. & Arneodo, A. (2016). Evidence of selection for an accessible nucleosomal array in human. *BMC Genomics* 17.

[8] Valouev, A., Johnson, S. M., Boyd, S. D., Smith, C. L., Fire, A. Z. & Sidow, A. (2011). Determinants of nucleosome organization in primary human cells. *Nature* 474, 516–520.

[9] Schones, D. E., Cui, K., Cuddapah, S., Roh, T.-Y., Barski, A., Wang, Z., Wei, G. & Zhao, K. (2008). Dynamic Regulation of Nucleosome Positioning in the Human Genome. *Cell* 132, 887–898.

[10] Brunet, F. G., Audit, B., Drillon, G., Argoul, F., Volff, J.-N. & Arneodo, A. (2018). Evidence for DNA Sequence Encoding of an Accessible Nucleosomal Array across Vertebrates. *Biophysical Journal* .

[11] Chalopin, D., Naville, M., Plard, F., Galiana, D. & Volff, J.-N. (2015). Comparative analysis of transposable elements highlights mobilome diversity and evolution in vertebrates. *Genome Biol. Evol.* 7, 567–580.

[12] Orgel, L. E. & Crick, F. H. (1980). Selfish DNA: the ultimate parasite. *Nature* 284, 604–607.

[13] McClintock, B. (1956). Controlling elements and the gene. *Cold Spring Harb. Symp. Quant. Biol.* 21, 197–216.

[14] Coyne, J. A. & Orr, H. A. (1998). The evolutionary genetics of speciation. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* 353, 287–305.

[15] Kraaijeveld, K. (2010). Genome size and species diversification. *Evol. Biol.* 37, 227–233.

[16] Valouev, A., Johnson, S. M., Boyd, S. D., Smith, C. L., Fire, A. Z. & Sidow, A. (2011). Determinants of nucleosome organization in primary human cells. *Nature* 474, 516–520.

[17] Pryciak, P. M. & Varmus, H. E. (1992). Nucleosomes, DNA-binding proteins, and DNA sequence modulate retroviral integration target site selection. *Cell* 69, 769–780.

[18] Naughtin, M., Haftek-Terreau, Z., Xavier, J., Meyer, S., Silvain, M., Jaszczyszyn, Y., Levy, N., Miele, V., Benleulmi, M. S. *et al.* (2015). DNA physical properties and nucleosome positions are major determinants of HIV-1 integrase selectivity. *PLoS One* 10, e0129427.

[19] Cournac, A., Koszul, R. & Mozziconacci, J. (2016). The 3D folding of metazoan genomes correlates with the association of similar repetitive elements. *Nucleic Acids Res* 44, 245–255.

[20] Mighell, A. J., Markham, A. F. & Robinson, P. A. (1997). Alu sequences. *FEBS Lett.* 417, 1–5.

[21] Deininger, P. (2011). Alu elements: know the SINEs. *Genome biology* 12, 236.

[22] Tang, W., Mun, S., Joshi, A., Han, K. & Liang, P. (2018). Mobile elements contribute to the uniqueness of human genome with 15,000 human-specific insertions and 14 Mbp sequence increase. *DNA Res.* 25, 521–533.

[23] Tang, W. & Liang, P. (2019). Comparative Genomics Analysis Reveals High Levels of Differential Retrotransposition among Primates from the Hominidae and the Cercopithecidae Families. *Genome Biol Evol* 11, 3309–3325.

[24] Durbin, R. M., Altshuler, D., Durbin, R. M., Abecasis, G. R., Bentley, D. R., Chakravarti, A., Clark, A. G., Collins, F. S., De La Vega, F. M. *et al.* (2010). A map of human genome variation from population-scale sequencing. *Nature* 467, 1061–1073. Number: 7319 Publisher: Nature Publishing Group.

[25] Auton, A., Abecasis, G. R., Altshuler, D. M., Durbin, R. M., Abecasis, G. R., Bentley, D. R., Chakravarti, A., Clark, A. G., Donnelly, P. *et al.* (2015). A global reference for human genetic variation. *Nature* 526, 68–74. Number: 7571 Publisher: Nature Publishing Group.

[26] Roy-Engel, A. M., Salem, A.-H., Oyeniran, O. O., Deininger, L., Hedges, D. J., Kilroy, G. E., Batzer, M. A. & Deininger, P. L. (2002). Active Alu element "A-tails": size does matter. *Genome Res* 12, 1333–1344.

[27] Drillon, G., Audit, B., Argoul, F. & Arneodo, A. (2015). Ubiquitous human 'master' origins of replication are encoded in the DNA sequence via a local enrichment in nucleosome excluding energy barriers. *Journal of Physics: Condensed Matter* 27, 064102.

[28] Arneodo, A., Drillon, G., Argoul, F. & Audit, B. (2018). The role of nucleosome positioning in genome function and evolution. In *Nuclear Architecture and Dynamics, Vol. 2* (C. Lavelle & J.-M. Victor, eds.), pp. 41–79. Elsevier.

Session 5, Amphi C

Knowledge representation, databases & visualization II

Discovery of potential functional paths by integration of phospho-proteomics data in the PPI network using a RWR framework

Jérémie PERRIN[1], Olivier DESTAING[2] and Christine BRUN[1,3]

[1] Aix-Marseille University, INSERM, TAGC, Turing Centre for Living Systems, 163 Avenue de Luminy, Marseille 13009, France

[2] Institute for Advanced Biosciences, Centre de Recherche Université Grenoble Alpes, Inserm U 1209, CNRS UMR 5309, 38706 La Tronche, France

[3] CNRS, 31 Chemin Joseph Aiguier, Marseille 13009, France

Corresponding author: jeremie.perrin@univ-amu.fr

Abstract *Understanding how cellular signalling is flowing from the molecular to the cellular level is a key step to identify regulators of different diseases and revisit the development of new potential drug targets. For years, biological approaches of signalling did not allow to probe and control signalling at the sub-cellular level with enough accuracy in space and time to directly witness transfer of information in biological network. To analyze datasets where signaling is controlled spatio-temporally by optogenetic, we have developed a method which traverses the space of Random Walks with Restart (*RWR*) models, searching for the optimally biased walk in a given context. It will allow to integrate data of differentially phosphorylated proteins obtained from longitudinal phospho-proteomics assay, in response to two different mode of optogenetic activation of the kinase Src, in order to reconstruct potential functional paths in the Protein-Protein interaction (*PPI*) network.*

Keywords Random Walk with Restart, optimization, PPI network, path finding, signalling

1 Introduction

Activation of a single intracellular signaling element can induce a decision making event: different mode of activation of the same cell can have very different phenotypic responses. This suggests that some mechanism downstream of the stimulus drives the signalling processes into two signalling directions, inducing two different cellular responses. In an optogenetic engineered system, Kerjouan et al. [1] show such phenomenon at play. They successfully construct a functioning photo-activable version of the Src tyrosine kinase. By either being able to restrict the kinase movements to the membrane surface (2D diffusion) or being able to let it diffuse freely inside the cytosol (3D diffusion), they manage to activate the same level of Src in seemingly similar situations. They show that slightly modulating diffusion of these signals in the same site of action is sufficient to induce very distinct cellular phenotypes. In the case of membrane diffusion the cells exhibit lamellipodia, whereas they exhibit invadosome structures when the kinase is not restrained.

In order to understand the mechanisms at play, we need to be able to reconstruct the signal transduction after the activation of the Src kinase. Since our ability in monitoring the phosphorylation statuses of all proteins in a cell is fairly limited, the phospho-proteomics approaches are costly and their ability to be precisely quantitative is questionable, our hope is to be able to reconstruct some potential transduction paths from a coarse time-resolved phospho-proteomic assay, leveraging the information contained in the topology of the PPI network. We develop our method in order to decipher the events occurring between the time of the optogenetic Src activation and the cellular responses observed. A visual overview of the method is given in Figure 1.

2 Methods

To reconstruct potential functional paths, we observe the change in phosphorylation levels of a wide range of proteins following the activation of the optoSrc (OS-sensitive proteins). We use the RWR paradigm (see section 2.1) to analytically compute how much the Src kinase is able to influence the rest of the PPI network. The ordinary RWR makes a strong assumption on the possible interactions

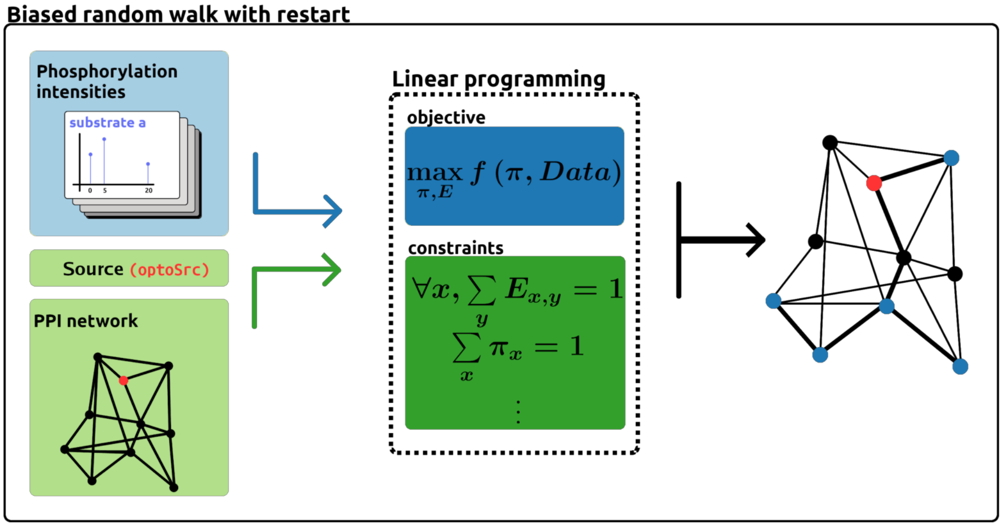

Biased random walk with restart
Phosphorylation intensities
substrate a
0
5
20
Source (optoSrc)
PPI network
Linear programming
objective
$\max_{\pi,E} f(\pi, Data)$
constraints
$\forall x, \sum_y E_{x,y} = 1$
$\sum_x \pi_x = 1$
⋮

Fig. 1. *(left)* Inputs of the algorithm: phosphorylation intensities, topology of the PPI network and the protein of interest *(center)* Formulation of the RWR problem as a linear program *(right)* Output of the algorithm: the biases of the optimal RWR model which we will analyse in downstream workflow.

a protein might have at a given time : it supposes that a given protein has uniform probability of interacting with any of its partners, leading to a sphere of influence which is not context dependent. In our setting we have the information that, after a certain time, the activation of our initial kinase led to specific changes in phosphorylation levels. We will use this knowledge to guide our random walk and deduce some potential contextual affinities between protein partners. By exactly describing the space of RWR models using linear constraints on both the edge weights and the asymptotic distribution (see section 2.2), as was done in [2] for random walks, we are able to use cutting-edge optimization software [3] to find the mathematically optimal edge weights for a given objective function. The objective function we will be interested in is the quantitative matching of the RWR's asymptotic distribution to the experimental observation (see section 2.4). We therefore retrieve the edge weights corresponding to a stable distribution which matches best our observations of phosphorylation levels.

2.1 Random Walk with Restart

Given a graph $\mathcal{G} = (V, E)$, where V is the set of vertices (nodes) of the graph and E the set of edges, a set of initial nodes $\mathcal{I}$ with associated vector R ($\forall i \in I, R_i = \frac{1}{|\mathcal{I}|}$ and $\forall i \in V, i \notin \mathcal{I} \implies R_i = 0$). We can define A the adjacency matrix of the graph. We will start by defining the set of matrices we are interested in :

Definition 2.1 (Stochastic matrices with support A).

$$\mathcal{S}_A := \left\{ M \in M_n(\mathbb{R}) \quad s.t \quad \forall j, \sum_i M_{i,j} = 1 \ \textit{and} \ \forall i, j, A_{i,j} = 0 \implies M_{j,i} = 0 \right\} \qquad (1)$$

We define $\mathcal{S}_A$ the set of stochastic matrices which have A as a support. These matrices are stochastic because they represent a Markovian process : the random walk. This random walk is restricted to the edges of our underlying graph that is why the matrices should have the same support as A. This is a description of the space of parameters for our random walks.

For any matrix $W \in \mathcal{S}_A$, the Random Walk with Restart using the weights W is defined as the process:

$$P_{t+1} = \beta R + (1 - \beta) W P_t$$

Ergodic theory (using the Perron-Frobenius theorem) shows that such a process converges. We will call the asymptotic distribution of the process $\mathrm{RWR}_\beta(W)$. If $\Pi = \mathrm{RWR}_\beta(W)$ then it satisfies :

$$\Pi = \beta R + (1 - \beta) W \Pi \quad \text{and} \quad \Pi = \sum_{k=0}^{\infty} \beta (1 - \beta)^k W^k R$$

In most cases, when there is no information on the weights, W is chosen to be uniform at each nodes *i.e* $W = A^T D^{-1}$ where D is the diagonal matrix of degrees, we will be referring to this choice as the ordinary RWR.

The Random Walk with Restart (*i.e* Network Diffusion, Personalized PageRank) is a classical procedure which computes similarity scores between nodes in a graph. It is often used in network biology, and network science in general, to either : determine proximity between pairs of nodes in biological networks [4] and more complex networks [5], assess quality of general clustering methods [6] and to predict directionality in undirected PPI networks [7].

The PPI network constructed from Y2H high throughput experiments does not contain information about affinity of interactions nor context dependent competition between proteins. Although the RWR can leverage edge weights in weighted networks, due to the lack of information, the choice is often made to use uniform weights but it is a strong assumption. We have devised a procedure to optimize the weights in order to match some observations on the network. The choice of the network that we will actually use in the context of the Src signalling will not be discussed here but we can point out that it is one of the educated choices which has to be made in order to have the best possible predictions.

2.2 Optimally biased RWR

If we have some observations on a subset $\mathcal{O} \subset V$ of the vertices at two different time points $\{t_0, t_1\}$ (*i.e* we have a function $obs : \mathcal{O} \times \{t_0, t_1\} \to \mathbb{R}^+$). We then are interested in finding the optimal parameters which explain the change in the observations.

Definition 2.2 (Set of pairs of matrices with their associated asymptotic distribution). *To every matrix we can associate a unique asymptotic distribution under the biased random walk with restart with parameters ϵ and β.*

$$\mathbb{RWR}_{\epsilon,\beta} \coloneqq \{(S, RWR_\beta((1-\epsilon)W_0 + \epsilon S))\,, \quad S \in \mathcal{S}_A\} \tag{2}$$

Instead of searching the whole space of stochastic matrices which have A as a support $\mathcal{S}_A$ (see Equation 1), which could give very unpractical results because it would not take into account diffusion at all, we will search the space of parameters around an *a priori* set of parameters W_0. The choices of W_0 could be one of the following: uniform probabilities, maximum entropy probabilities, optimized probabilities to match the initial observation (the actual choice will not be discussed here but is a crucial point to be studied in the future). To summarize, we will be considering, for a given ϵ, all matrices of the form $(1-\epsilon)W_0 + \epsilon S$ for S in $\mathcal{S}_A$, those are the biased random walks around W_0 (see Equation 2). Out of all of these matrices we will try to find the one which best explains the changes in the observations. If we have a function f to compare the asymptotic distribution of the RWR to the observations then we are interested in :

$$S^* \in \underset{S,\Pi \in \mathbb{RWR}_{\epsilon,\beta}}{\operatorname{argmin}} \; f\,(\Pi, obs) \tag{3}$$

Remark 2.3. *In our case, the observations will be the phosphorylation intensities measured for each protein for which we actually have a value, before and after Src activation. We will discuss the actual choice of f in section 2.4, but first let us describe the space of all possible RWRs.*

Remark 2.4. *These optimization problems are hard in general, except when the problem can be formulated in certain ways. This is the theory of convex optimization. We will not address the mathematics underlying the optimization procedures but we will show that our problem can be formulated as a Linear Program. Linear Programs are a kind of well studied optimization problems for which we have good enough optimizers for problems of the size we are interested in.*

2.3 Specifying the convex search space

We want to describe the set $\mathbb{RWR}_{\epsilon,\beta}$, the stochastic matrices with support matrix A as well as their associated asymptotic distribution for the Random Walk with Restart. We will give the linear constraints which describe this set as a subspace of $\mathbb{R}^{|V|^2} \times \mathbb{R}^{|V|}$.

Let us start with $\mathcal{S}_A$:

$$S \in \mathcal{S}_A \iff \forall j, \quad \sum_{i=0}^{|V|} S_{ij} = 1 \text{ and } \forall ij, \quad 0 \leq S_{ij} \leq A_{ji} \tag{4}$$

Now let us describe $\text{RWR}_\beta((1-\epsilon)W_0 + \epsilon S)$:

$$\Pi = \text{RWR}_\beta((1-\epsilon)W_0 + \epsilon S) \iff \quad \sum_j \Pi_j = 1 \tag{5}$$

$$\text{and} \quad \forall j, \quad 0 \leq \Pi_j \tag{6}$$

$$\text{and} \quad \Pi = \beta R + (1-\epsilon)(1-\beta)W_0\Pi + \epsilon(1-\beta)S\Pi \tag{7}$$

Here Equation 7 is not linear, the problem is solved by a simple change of variables $E \coloneqq S \times \text{diag}(\Pi)$ and rewriting the equations subsequently (similarly to what is done in [2]):

$$\forall j, \quad \sum_{i=0}^{|V|} E_{ij} = \Pi_j \text{ and } \forall ij, \quad 0 \leq E_{ij} \leq A_{ji} \tag{4bis}$$

$$\Pi = \beta R + (1-\epsilon)(1-\beta)W_0\Pi + \epsilon(1-\beta)E \tag{7bis}$$

2.4 Objective function

Let us discuss the choice of f, in order for the problem to be a Linear Program, we need f to be linear in both E and Π. Since our goal is to match the observations, it actually will not depend on E (although we could later add some regularization term to our objective which could depend on E). In fact we already hinted that way in Equation 3, by not having E be an argument of f.

Our goal is to find the edge weights which bias the random walk's asymptotic distribution towards the activated nodes and away from the inactivated nodes. Thus our first approach was to consider the sets of nodes $\mathcal{O}^+ \coloneqq \{i \in \mathcal{O}, \quad obs(t_1) - obs(t_0) > 0\}$ and $\mathcal{O}^- \coloneqq \{i \in \mathcal{O}, \quad obs(t_1) - obs(t_0) < 0\}$. Trying to maximize probabilities of ending on nodes in $\mathcal{O}^+$ which have a positive variation of the observed quantity and minimize probabilities of ending on nodes in $\mathcal{O}^-$ which have negative variation of observed quantity :

$$f_1(\Pi, obs) = -\left(\sum_{i\in\mathcal{O}^+} \Pi_i - \sum_{i\in\mathcal{O}^-} \Pi_i\right)$$

By not taking into account the actual structure of RWR (*i.e* correlation between distance from the source and the probability of ending on the node), this tends to try and optimize the probabilities on nodes that are closest to the initial nodes. We solved this issue by looking at the fold change in probability from the initial condition $\text{RWR}(W_0)$:

$$f_2(\Pi, obs) = -\left(\sum_{i\in\mathcal{O}^+} \frac{\Pi_i}{\text{RWR}(W_0)_i} - \sum_{i\in\mathcal{O}^-} \frac{\Pi_i}{\text{RWR}(W_0)_i}\right)$$

Now this was satisfactory in practice, but it didn't leverage the whole information of the observations: we only considered the trend of the observed quantity, not the actual values. We would rather be matching the actual fold change in the observed quantity :

$$f_3(\Pi, obs) = \sum_{i\in\mathcal{O}} \left| \frac{\Pi_i}{\text{RWR}(W_0)_i} - \frac{obs(i,t_1)}{obs(i,t_0)} \right|$$

The absolute values are not linear, but there exists a trick which introduces new variables in the Linear Program to transform absolute values in the objective function into constraints :

$$\begin{aligned} \text{Minimizing } f_3 \iff \quad & \text{Minimizing } f_4 = \sum_{i \in \mathcal{O}} X_i \\ \text{and} \quad & X_i \geq \frac{\Pi_i}{\mathrm{RWR}(W_0)_i} - \frac{obs(i, t_1)}{obs(i, t_0)} \\ \text{and} \quad & X_i \geq - \left(\frac{\Pi_i}{\mathrm{RWR}(W_0)_i} - \frac{obs(i, t_1)}{obs(i, t_0)} \right) \end{aligned}$$

2.5 Solving

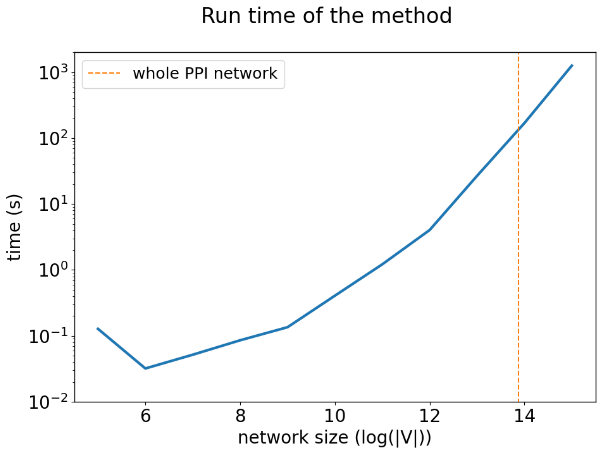


Fig. 2. Run times of the method on synthetic data. We have generated networks of different sizes, using Watts–Strogatz model. We have kept the ratio of number of edges and number of vertices fixed in all generated graphs. We consistently used a single source and $\frac{|V|}{30}$ target nodes.

Now that we have specified our problem as a Linear Program in a satisfactory way, we can compute the optimal solution *i.e* the optimal deviation from the initial condition to match our observations. If we come back to our biological question, we have a description of those interactions which are potentially favored or unfavored, in the context of our observations.

We are using the gurobi optimizer in python, through the gurobipy package. We have an academic license, but the software is proprietary. In order to support open source ecosystems, we might use the CLP software [8] in the future.

Remark 2.5 (Scalability). *Scalability does not seem to be an issue, see Figure 2, although restricting the network to functional modules will definitely run faster than on the whole PPI network.*

3 Results

As for now, we have tested our method on synthetic data. A visual representation of the output of the method on a 6 × 6 grid with single source and two positive targets is given in Figure 3.

From the output of our method we are capable of extracting paths from sources to targets, either from the probability landscape (Figure 3.A) or from the bias of our optimal RWR (Figure 3.B). In the actual workflow on the Src activation problem, we will develop some further downstream analysis to study the paths. In the Src context we will actually get a couple of paths we will need to compare in order to determine proteins/interactions which are potentially responsible for the decision making. We will also discuss validation of the method on gold standard data in section 4.

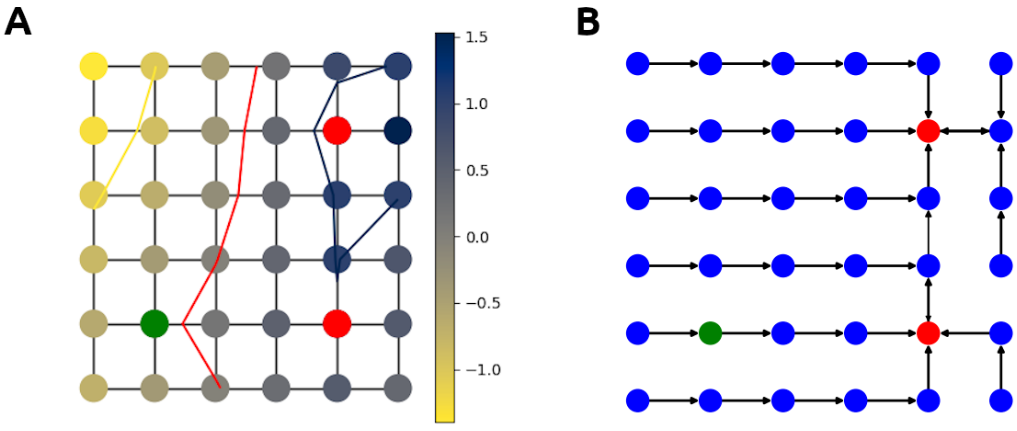


Fig. 3. (A & B) Green node is the source node, both red targets have a positive variation in the observed quantity **(A)** Probability landscape as a $\log_2$ fold-change between $\mathrm{RWR}(W_0)$ and $\mathrm{RWR}((1-\epsilon)W_0 + \epsilon S^*)$ **(B)** Representation of S^* as a directed graph, we visualize the biases *i.e* the directions favored in order to optimize the probabilities on the target nodes.

4 Discussion

We have presented our method, some technical details and some choices we have made based on our understanding of both the mathematical tool we are using and the biological context in which we want to apply the method. This being a work in progress, a substantial amount of leeway remains. One could question the biological relevance of the paths we find, and would be legitimate in doing so. We do not pretend to be providing a model of signalling, the potential functional path we are discovering will help us guide new experiments and question the roles of specific proteins, but should not be seen as predicted signalling pathways. In order to reinforce our trust that there is biological significance in the paths we discover, we will have to validate our method and adapt it to integrate other sources of biological information.

There are a couple of instances where we can, and will, try to validate our approach. The first case is in trying to reproduce directionality information as was done in [7]: we can try to predict the direction of interactions and confront our predictions to the ground truth since some interactions are known to be directed. Another idea is in trying to retrieve signalling pathways, from an obfuscated version of said pathways. Indeed, we could consider a signalling pathway, forget the directionality of the interactions, add some partner proteins which are not part of the pathway (obfuscating the original pathway) and then see if our method is capable of retrieving some parts of the pathway. Seeing how the method fares in controllable settings, will define how relevant we consider the paths discovered *de novo*.

A criticism regarding the interpretability of the RWR in the context of the PPI network comes from the fact that the PPI is constituted of binary interactions which are tested in conditions very different from the condition in which the interactions actually occur. The ”real” network representation of the protein interaction at a given time in a given cell has to differ from the PPI network (some interactions might be context-dependant). This observation drives our work in the direction of integrating other sources of information into the RWR paradigm. We are currently working on different ways of integrating orthogonal data to the Random Walk with Restart (proteomics, functional annotation). We have already mentioned that the choice of the underlying network is crucial, we have already built sub-networks of the PPI from proteins extracted from the literature, and are thinking about integrating data from Phosphosite+ [9]. We have developed our method in order to be able to integrate cause-to-effect contextual information, adding even more biological information should bring more specific results.

Acknowledgements

This work is part of Jérémie Perrin's PhD thesis. We thank Lionel Spinelli and Saswati Saha for their useful advices and precious discussions.

References

[1] Kerjouan *et al.* Control of SRC molecular dynamics encodes distinct cytoskeletal responses by specifying signaling pathway usage. *Journal of Cell Science*, 134(2), 01 2021.

[2] Yossi Azar, Andrei Z Broder, Anna R Karlin, Nathan Linial, and Steven Phillips. Biased random walks. *Combinatorica*, 16(1):1–18, 1996.

[3] Gurobi Optimization, LLC. Gurobi Optimizer Reference Manual, 2022.

[4] Hao-xiang Qi, Qi-dong Shen, Hong-yi Zhao, Guo-zhen Qi, and Lei Gao. Network-based analysis revealed significant interactions between risk genes of severe covid-19 and host genes interacted with sars-cov-2 proteins. *Briefings in Bioinformatics*, 23(1):bbab372, 2022.

[5] Alberto Valdeolivas, Laurent Tichit, Claire Navarro, Sophie Perrin, Gaelle Odelin, Nicolas Levy, Pierre Cau, Elisabeth Remy, and Anaïs Baudot. Random walk with restart on multiplex and heterogeneous biological networks. *Bioinformatics*, 35(3):497–505, 2019.

[6] Alexander J Gates, Ian B Wood, William P Hetrick, and Yong-Yeol Ahn. Element-centric clustering comparison unifies overlaps and hierarchy. *Scientific reports*, 9(1):1–13, 2019.

[7] Dana Silverbush and Roded Sharan. A systematic approach to orient the human protein–protein interaction network. *Nature communications*, 10(1):1–9, 2019.

[8] Robin Lougee-Heimer. The common optimization interface for operations research: Promoting open-source software in the operations research community. *IBM Journal of Research and Development*, 47(1):57–66, 2003.

[9] Peter V Hornbeck, Bin Zhang, Beth Murray, Jon M Kornhauser, Vaughan Latham, and Elzbieta Skrzypek. Phosphositeplus, 2014: mutations, ptms and recalibrations. *Nucleic acids research*, 43(D1):D512–D520, 2015.

Semantic Web technologies are effective to remove redundancies from protein-protein interaction databases and generate reproducible interactomes

Marc MELKONIAN[1,2], Camille JUIGNÉ[1,3], Olivier DAMERON[1], Gwenaël RABUT[2] and Emmanuelle BECKER[1]

[1] Univ Rennes, Inria, CNRS, IRISA, Rennes, F-35000, France.
[2] Univ Rennes, CNRS, IGDR-UMR 6290, Rennes, F-35000, France.
[3] Pegase, Inrae, Institut Agro, Saint-Gilles, 35590, France.

Corresponding author: marc.melkonian@irisa.fr

Reference paper:* Melkonian *et al.* (2022) Towards a reproducible interactome: semantic-based detection of redundancies to unify protein-protein interaction databases. *Bioinformatics. http://dx.doi.org/10.1093/bioinformatics/btac013

Information on protein-protein interactions (PPIs) is collected in numerous primary databases with their own curation process. To provide more exhaustive datasets, several meta-databases aggregate PPIs from multiple primary databases. However, aggregation of PPIs from different primary databases is not straightforward since distinct databases are often partly redundant and may have different PPI annotation policies. Mere aggregation can thus introduce a bias if these redundancies are not identified and eliminated, leading to systematically overestimating PPI reproducibility.

We propose a precise definition of two types of redundancies that can be observed between entries of PPI databases. We define explicit redundancy as the exact duplication of a PPI information in two distinct database entries. It occurs in particular when two databases have independently registered the same experimental evidence from a given publication and annotated it using an identical interaction detection method (IDM) term from the Proteomics Standards Initiative-Molecular Interactions (PSI-MI) ontology. Explicit redundancy is trivial and usually eliminated during the aggregation process of meta-databases. In contrast, we define implicit redundancy as the occurrence of two database entries referring to the same PPI reported in a given publication, but annotated with related IDM terms. Implicit redundancy occurs in particular when primary databases have different PPI annotation policies and preferentially use different types of IDM terms (for instance general vs specific). We show that both types of redundancies can be easily detected using Semantic Web technologies.

Using a dataset from the APID meta-database, we observed that while explicit redundancies are detected by the APID aggregation process, about 15% of APID entries are implicitly redundant. More than 90% of these redundancies result from the aggregation of distinct primary databases (inter-database redundancy), while the remaining occurs between entries of a single database (intra-database redundancy).

Further, we built for two species (yeast and human) a ”reproducible interactome” with interactions that have been reproduced by multiple methods or publications. For both species, the size of the reproducible interactome is drastically impacted by removing redundancies (-59% and -56% for yeast and human, respectively), and we show that this is largely due to implicit redundancies. Thus, a significant number of PPIs currently considered as reproducible actually relies on database integration artefacts.

All software are are freely available at https://gitlab.com/nnet56/reproducible-interactome and data and results can be browsed and downloaded at https://reproducible-interactome.genouest.org/.

Investigating ADR mechanisms with Explainable AI: a feasibility study with knowledge graph mining

Emmanuel Bresso[1,2], Pierre Monnin[1,2], Cédric Bousquet[2], François-Elie Calvier[2], Ndeye-Coumba Ndiaye[2], Nadine Petitpain[2], Malika Smaïl-Tabbone[1] and Adrien Coulet[1,2]

[1] Université de Lorraine, CNRS, Inria, LORIA, Nancy, France

[2] Please see [1] for full author information

Corresponding author: adrien.coulet@inria.fr

Reference paper: **Bresso *et al.* (2021) Investigating ADR mechanisms with Explainable AI: a feasibility study with knowledge graph mining. *BMC Medical Informatics and Decision Making.*** https://doi.org/10.1186/s12911-021-01518-6

Adverse drug reactions (ADRs) are statistically characterized within randomized trials or by post-marketing pharmacovigilance. However, the molecular mechanisms causing ADRs remain unknown in most cases. This is true even for common toxicities that are classically monitored during trials such as hepatic or skin toxicities. Interestingly, many elements of knowledge about drugs and drug ingredients are available beside clinical trials. In particular, open-access knowledge graphs describe their properties, interactions, and involvements in pathways. Expert classifications have also been manually established by experts and label drugs either as causative or not for several types of ADRs.

In our paper [1], we propose to mine biomedical knowledge graphs to identify biomolecular features that enable to automatically reproduce such expert classifications, distinguishing drugs causative or not for a given type of ADR. In an Explainable AI perspective, we explore simple classification techniques such as Decision Trees and Classification Rules because they provide human-readable models which explain the classification itself. We also evaluate the assumption that biomolecular features mined from knowledge graphs might provide elements of explanation for the molecular mechanisms behind ADRs. We tested our approach with two expert classifications identifying drugs causing or not hepatic or skin toxicities (respectively named *DILI* and *SCAR* for Drug Induced Liver Injuries and Severe Cutaneous Adverse Reactions). Features associated with these drugs were mined from PGxLOD [2], a biomedical knowledge graph that we previously created by interlinking public open data (including DisGeNET, PharmGKB, DrugBank, CTD). To this aim, we developed the `kgpm` algorithm [3] that enables scaling the extraction of paths of features up to a length of 4. Such paths are subsequently generalized into path patterns to cover larger sets of drugs. We trained two classifiers at distinguishing, on the basis of extracted features, drugs causing or not each of the two considered ADRs. We isolated features that are both discriminative in reproducing expert classifications and interpretable by experts (*e.g.*, *Gene Ontology* terms, drug targets, pathways) and asked 3 pharmacology experts to manually evaluate if they are potentially explanatory for ADRs.

Extracted features reproduce DILI and SCAR classifications with a fair fidelity (accuracy of 0.74 and 0.81, respectively). Experts fully agreed that 73% and 38% of the most discriminative features are possibly explanatory for DILI and SCAR, respectively; and partially agreed (2/3) for 90% and 77% of them. To illustrate, the path pattern $\xrightarrow{\text{interactsWith}}$ Enzyme $\xrightarrow{\text{cellularComponent}}$ Endoplasmic reticulum reached an agreement for DILI. Indeed, the endoplasmic reticulum is known, in particular in liver tissues, to host primarily cytochrome P450 enzymes, well known for being involved in drug metabolism.

From these results, it appears that the considered knowledge graph provides sufficiently diverse features to enable simple and explainable models to distinguish between drugs that are causative or not for ADRs. In addition to explaining expert classifications, most discriminative features appear to be good candidates for investigating further molecular mechanisms behind ADRs.

References

[1] Emmanuel 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.

[2] Pierre 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] Pierre Monnin et al. Tackling scalability issues in mining path patterns from knowledge graphs: a preliminary study. In *First International Conference "Algebras, graphs and ordered sets", ALGOS*, 2020.

Symposium 1, Amphi A

Indexing and querying of sequencing data

Indexing and querying of sequencing data

Organizers: Pierre PETERLONGO[1], Camille MARCHET[2], Rayan CHIKHI[3], Antoine LIMASSET[2], Clément AGRET[4], Téo LEMANE[1], Lucas ROBIDOU[1] and Karel BRINDA[1]

[1] Univ Rennes, CNRS, Inria, IRISA-UMR 6074, F-35000 Rennes, France

[2] Univ. Lille, CNRS, UMR 9189 - CRIStAL, F-59000 Lille, France

[3] Department of Computational Biology, C3BI USR 3756 CNRS, Institut Pasteur, Paris, France

[4] ETIS UMR 8051, CY Paris Cergy University, ENSEA, CNRS, F95000 Cergy, France

This mini-symposium is supported by:

cnrs GDR Groupement de recherche
BIM Bioinformatique moléculaire

Data-structures for querying large k-mer (collections of) sets

Camille MARCHET

Univ. Lille, CNRS, UMR 9189 - CRIStAL, F-59000 Lille, France

Invited speaker: camille.marchet@univ-lille.fr

High-throughput sequencing datasets are usually deposited in public repositories, e.g. the European Nucleotide Archive, to ensure reproducibility. As the amount of data has reached petabyte scale, repositories do not allow to perform online sequence searches; yet such a feature would be highly useful to investigators. Towards this goal, in the last few years several computational approaches have been introduced to index and query large collections of datasets. In this seminar I propose an overview of methods for representing and indexing sets of k-mer efficiently. Then we will review how these techniques were adapted to index collections of thousands of datasets (and more) for membership queries. I will propose application examples for these techniques with a focus on RNA and splicing.

Scalable Indexing of Sequencing Data in Annotated Genome Graphs

Mikhail KARASIKOV

ETH Zurich, Department of Computer Science, Biomedical Informatics Group, Zurich, Switzerland; University Hospital Zurich, Biomedical Informatics Research, Zurich, Switzerland; Swiss Institute of Bioinformatics, Lausanne, Switzerland

Invited speaker: mikhaika@inf.ethz.ch

Technological advances have led to an exponential growth in the amount of high-throughput sequencing data available to the scientific community. However, most of this data, commonly stored in repositories such as the NCBI Sequence Read Archive, which currently contains over 50 petabases of sequences, is only available in raw format. Transforming it into a searchable representation that is easily accessible to life science researchers for large scale analysis and search remains an unsolved problem.

In my talk, I will review the state-of-the-art approaches for indexing large cohorts of sequencing data. Then, I will describe MetaGraph, a method that efficiently indexes petabase-scale cohorts of sequencing experiments in annotated de Bruijn graphs and that supports k-mer search and sequence-to-graph alignment. Internally, MetaGraph represents input data as collections of k-mer sets encoded in succinct data structures, offering practically relevant trade-offs between the index size and the query performance. This flexibility allows running MetaGraph at different scalesand on different hardware, from laptops to research compute clusters and distributed cloud environments.

I will pay special attention to the methods and data structures used in MetaGraph for representing graph annotations, including non-binary attributes with Counting de Bruijn graphs, e.g., for representing gene expression and genome coordinates. Finally, I will conclude with real-world applications, such as indexing a portion of all publicly available whole-genome sequencing samples from the Sequence Read Archive, currently including over 90% of all Microbe, Fungi, Plant, and Human, as well as indexing all reference genome sequences (RefSeq), the RNA-Seq Genotype-Tissue Expression dataset (GTEx), and significantly more diverse metagenomic data, such as the entire catalog of 286,997 reference genome sequences from the human gut microbiome (UHGG), all 242,619 publicly available human gut microbiome short read sequencing samples, and a set of 4,220 public transit surface microbiomes (MetaSUB).

TranSipedia: a novel framework for large scale RNAseq data analysis with applications in cancer from research to diagnosis

Thérèse COMMES

IRMB, University of Montpellier, INSERM, 80 rue Augustin Fliche,
34295, Montpellier, France

Invited speaker: therese.commes@inserm.fr

Driven by myriads of projects, public RNAseq databases are exploding, to date, over 850,000 RNA-seq are deposited on SRA for humans. This huge body of publicly available RNAseq libraries is a precious resource to identify specific transcriptional events. The challenges lie in the complexity of RNA biological content and the exponential increase in data volume. We want to make RNA-seq data easily accessible providing a better capture of the whole transcriptome complexity, in the context of biological and human health applications. New computational methods that perform indexing of k-mers across huge datasets constitute interesting solutions to interrogate "Omics data" at a large scale from dataset collections. Here, we developed TranSipedia, a new framework, based on kmer approach, constructed with several modules: 1/ The RNA-seq indexing step constructed with Reindeer (REad Index for abuNDancE quERy; Marchet et al., 2020), a novel computational method that serves as an efficient platform to request all transcribed information, 2/a completed module to generate kmers as signature of transcripts (Kmerator; Riquier et al, 2021), 3/ a supporting web site to facilitate the queries easily shared by biologists (TranSipedia, https://transipedia.montp.inserm.fr).

Reindeer performs indexing of k-mers and records their counts across a large collection of datasets. Interestingly it associated k-mers to their counts instead of only recording the presence/absence of k-mers as frequently done in previous works. Moreover, Reindeer provides an ultra-fast performance in the query process while indexing several thousands of RNAseq. One of the great advantages of indexing raw data is also that it integrates reference-free and annotation-free approaches. For applications where gene expression level is required, the kmer count must be sufficiently sensitive and representative to be applicable. The quantitative accuracy with k-mers counts from Reindeer indexed datasets was compared to classical quantitative methods like Kallisto. Secondly, the design kmer module uses Kmerator, a tool develop to construct specific kmers already avalailable on github. Thirtly, the website is already available to facilitate index queries by the biologists with kmer on fasta file format. The Transipedia platform now deserves several thousands of datasets from public and private collections mainly from acute myeloid leukemia (AML) for cancer applications. We indexed whole CCLE cohort representing 1019 RNAseq samples for a total of 10To and f indexes from Leucegene, BEAT-AML, SRA and TCGA collections represented with more than 1000 RNAseq samples for AML application. Concerning biological applications, we already requested in selected public dataset gene co-expression, biomarker tissue specificity, as well as tumor specific signatures comparing normal/tumor, for simple and useful medical usage. In perspectives, based on data structures, such as k-mer features, diagnosis applications are in development. Moreover,Machine learning models could be used to search for signatures and explore better diagnostic and prognosis models.

The Tree of Life enables efficient and robust compression and search of microbes

Karel BRINDA

Univ Rennes, CNRS, Inria, IRISA-UMR 6074, F-35000, Rennes, France

Invited speaker: karel.brinda@inria.fr

Comprehensive collections of genomes have been instrumental in the study of life using sequencing; however, their storage, transmission, and analysis have become challenging due to the exponential growth of sequence data. This raises the question of designing efficient computational solutions for storing and indexing large data sets, such as the recently created comprehensive data set of 661k bacterial genomes, uniformly assembled from the European Nucleotide Archive.

Here, we present MOF (Microbes on a Flash Drive), a method for large-scale lossless compression and search of microbial collections, using the Tree of Life as a biological prior on their redundancy structure. MOF uses state-of-the-art tools and databases from population genomics and metagenomics to infer the geometrical structure of a given collection and identify the likely position of individual microbes on the Tree of Life, and then uses this information for guiding data compression using standard approaches.

We show MOF losslessly compresses 425 thousand microbial de Bruijn graphs from 16 TB (McCortex) to 52.2 GB (xz) (319×), 661 thousand bacterial assemblies from 805 GB (gzip) to 29 GB (xz) (27×), and their associated k-mer index from 937 GB (COBS) to 110 GB (xz) (8.5×).

When combined together as building blocks, this enables performing BLAST-like alignments to the 661k data set on a standard desktop computer within several hours, as we demonstrate with the examples of resistance genes, plasmids, and nanopore reads as queries. The optimization of data structures using the Tree of Life has broad applications across computational biology and provides a fundamental design principle for future genomics infrastructure.

Symposium 2, Amphi B

Management and integration of agronomic, phenotypic and environmental data

Management and integration of agronomical, phenotypical and environmental data

Organizers: Fabrice LEGEAI[1,2], Olivier DAMERON[2], Frédéric DE LAMOTTE[3], Yvan LE BRAS[4], Nicolas PARISEY[1] and Cyril POMMIER[5]

[1] INRA, UMR1349, Institute of Genetics, Environment and Plant Protection, 35653 Le Rheu, France

[2] Univ Rennes, CNRS, Inria, IRISA-UMR 6074, F-35000 Rennes, France

[3] UMR AGAP, INRAE, 34060 Montpellier, France

[4] Pôle National de Données de Biodiversité, UMS 2006 PatriNat, Station Marine de Concarneau, Muséum National d'Histoire Naturelle, F-29900 Concarneau, France

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

GDR MADICS, l'infrastructure PHENOME et le département BAP de l'INRAE

This mini-symposium is supported with the MADICS GDR, the PHENOME infrastructure, and the BAP department of INRAE.

Central role of metadata language to create a federated data endpoint with a high FAIRness degree

Yvan LE BRAS

Pôle National de Données de Biodiversité, UMS 2006 PatriNat, Station Marine de Concarneau, Muséum National d'Histoire Naturelle, F-29900 Concarneau, France

Invited speaker: yvan.le-bras@mnhn.fr

The “PNDB”, the French Biodiversity Data hub, is an e-infrastructure created in 2018 and entering an operational implementation phasis for these tools and services. Through the 3 components data/metadata, animation and data analysis, I propose to illustrate these tools and services by focusing on the approach proposed to go from data to the use of terminological resources via the intensive use of a metadata standard with multiple capabilities: the Ecological Metadata Language, or EML!

Use of deep learning methods applied to the stud of ticks and tick-borne diseases

Sk Imran HOSSAIN[1,2,3], Yann FRENDO[1,2,3], Valérie POUX[2,3], Isabelle LEBERT[2,3], Delphine MARTINEAU[4], Olivier LESENS[4,5], Engelbert MEPHU NGUIFO[1], Gwenaël VOURC'H[2,3] and Jocelyn DE GOËR DE HERVE[2,3]

[1] Université Clermont Auvergne, CNRS, ENSMSE, LIMOS, F-63000, Clermont-Ferrand, France

[2] Université Clermont Auvergne, INRAE, VetAgro Sup, UMR EPIA, F-63122, Saint-Genès-Champanelle, France

[3] Université de Lyon, INRAE, VetAgro Sup, UMR EPIA, F-69280, Marcy l'Etoile, France

[4] Infectious and Tropical Diseases Department, CHU Clermont-Ferrand, F-63000, Clermont-Ferrand, France

[5] CHU Clermont-Ferrand, Inserm, Neuro-Dol, CNRS 6023 Laboratoire Microorganismes Génome Environnement (LMGE), Université Clermont Auvergne, F-63000, Clermont-Ferrand, France

Invited speaker: jocelyn.degoer@inrae.fr

Use of deep learning methods applied to the study of ticks and tick-borne diseases. *Sk Imran* Ticks are important vectors of pathogens for human and animal populations in Europe. In a context of environmental changes, they are an Important source of concern for the population and a source of health risks that are frequently relayed by the public media. Therefore, surveillance and prevention are public health issues, which require the development of scientific research. These research projects are carried out in partnership with public research units, private partners or local collectivities. They involve observation systems (tick collection in the environment, observation and declaration of bites by citizens) and analysis of the data collected via statistical modelling, genomics or artificial intelligence methods.

In this context, we present two research projects illustrating the use of Artificial Intelligence (AI), in particular Deep Learning methods, that are applied to the problematics of ticks and Lyme disease:

The DAPPEM project (Development of an application for the Identification of Erythema Migrans) aimed to develop a smartphone application to detect the early signs of Lyme

disease, called erythema migrans, from a photo and a context questionnaire. The detection is performed using a Convolutional Neural Network (CNN) that automatically analyzes photos taken by users and returns a risk score calculated together with the context questionnaire. In the case of a suspected erythema migrans, a message will guide the user to an appropriate care pathway. The application is currently in the process of certification as a medical device. It will be available in 2023, to healthcare professionals and to the general public. The data that would be collected anonymously will be used to improve the detection model and to contribute to statistical studies for estimating incidence of Lyme disease.

The DCLIC project (Deep Convolutional Learning Ixodidae Characterization) aims to develop a tool that can detect in real time ticks in a video and classify them within four taxonomic genus that are present in metropolitan France (Ixodes, Hyalomma, Rhipicephalus, Dermacentor) The works carried out during this project have demonstrated that it was possible to automate the identification of tick genus from photos or videos. The latest version of the neural network is able to detect a tick from an image with an accuracy of 0.968 and a specificity of 0.963, when ticks are larger than 5% of the total surface of the photo. The methodological developments produced during the DCLIC project are intended to be shared with the scientific community and could be integrated into applications for research teams collecting ticks or into mobile applications for participatory action research.

« Sicpa IoT » : un déploiement de capteurs environnementaux, à large échelle, en unités expérimentales animales à INRAE

François LAPERRUQUE

CATI SICPA, INRAE, Toulouse INP, 31320, Castanet Tolosan, France

Orateur invité: francois.laperruque@inrae.fr

Le Cati SICPA (Systèmes d'Informations et Calcul pour le Phénotypage Animal) met en place des solutions de collecte, de transfert et de stockage des données, issues des unités expérimentales animales des départements Génétique Animale (GA) et Physiologie Animale et Systèmes d'Élevages. Sa composante calcul, contribue au déploiement de solutions pour l'analyse de données (infrastructure, logiciels, IA).

Dans le cadre du projet "Sicpa IoT", nous répondons aux demandes réglementaires et scientifiques d'enregistrements longitudinaux de données environnementales, en élevages. Les informations collectées sont importantes pour mieux comprendre l'expression des phénotypes des animaux (consommation d'eau, résistance aux maladies, comportement, ...). Le choix d'une infrastructure Lora, gérée en propre, a été réalisé. Celle-ci s'appuie sur un écosystème de standards ouverts et de logiciels libres, bénéficiant d'une large communauté d'utilisateurs.

Nous utilisons, par exemple, le serveur LoRaWAN Chirpstack, un broker MQTT pour transporter les données, ainsi que la base de données Time Series, InfluxDB pour leur stockage. L'ensemble de l'infrastructure s'appuie sur la solution OpenStack proposée par la DSI INRAE.

L'utilisation de standards ouverts va nous permettre de développer des interfaces simplifiées selon les types d'utilisateurs et de connecter des capteurs communicants à des formats divers.

PHIS and OpenSILEX information systems for FAIR data

Isabelle ALIC, Anne TIREAU, Arnaud CHARLEROY, Gabriel BESOMBES,
Valentin RIGOLLE and Pascal NEVEU
MISTEA, Univ Montpellier, INRAE, Montpellier SupAgro, 34060, Montpellier, France

Invited speakers: gabriel.besombes@inrae.fr, valentin.rigolle@inrae.fr

Research data are more and more complex and must evolve continuously to meet the scientific challenges of today and tomorrow. The information systems of research communities must therefore be able to handle this complexity and also meet the requirements of an open and reproducible science. In response to these challenges, we have developed OpenSILEX, a software suite that allows building information systems and implementing FAIR principles. These systems are ontology-driven. OpenSILEX is derived under several implementations corresponding to specific scientific communities:

PHIS (high throughput plant phenotyping),

Sixtine (experimental plant units at INRAE),

Silex-VitiOeno (wine and viticulture),

Silex-LBE (depollution).

The PHIS implementation is the information system of the Phenome project. It proposes, beyond data management, a structured vocabulary for a better data sharing within the community.

Rule Mining in Knowledge Graphs for Data integration and Beyond

Fatiha SAÏS

LRI, UMR 8623, CNRS, Université Paris-Sud, Université Paris Saclay,
F-91190, Gif-sur-Yvette, France

Invited speaker: fatiha.sais@lri.fr

Today, we are experiencing an unprecedented production of resources, published as Linked Open Data (LOD, for short). This is leading to the creation of knowledge graphs (KGs), such as DBpedia, YAGO and Wikidata on the academic side, and the Google Knowledge Graph or Microsoft's Satori graph on the commercial side. These KGs contain millions of entities (such as people, proteins, or books), and millions of facts about them. This knowledge is typically expressed in RDF (Resource Description Framework), i.e., as triples of the form ⟨Biden, presidentOf, USA⟩. Some KGs provide an ontology expressed in OWL2 (Web Ontology Language), which describes the vocabulary (the classes and properties) for the RDF facts. However, these knowledge graphs are far from being correct and complete. That is, the existing KGs need to be refined in order to be enriched while preventing introducing errors. In this talk, I will present some contributions for KG refinement task that we tackle through the mining of logical rules. These rules may be used for integrating heterogeneous information or to explain « causality » relationships between some phenomena. I will finish by highlighting some current challenges.

Symposium 3, Amphi C

Bioinformatics of metabolic pathways, from sequences to molecules

Mini-Symposium

Bioinformatics of metabolic pathways, from sequences to molecules

Organizers: Gabriel MARKOV[1], Samuel BERTRAND[2], Clément FRAINAY[3],
Clémence FRIOUX[4] and Jeanne GOT[5]

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

[2] Université de Nantes, Groupe Mer, Molécules, Santé-EA 2160, UFR des Sciences Pharmaceutiques et Biologiques, F-44035, Nantes, France

[3] INRA, UMR1331, Toxalim, F-31000 Toulouse, France

[4] Inria, INRAE, Université de Bordeaux, 33405, Talence, France

[5] Univ Rennes, CNRS, Inria, IRISA-UMR 6074, F-35000 Rennes, France

This mini-symposium is supported by the Inria centre at Rennes University, the Francophone Network of Metabolomics and Fluxomics, the MetaboHub infrastructure, the French Chemoinformatics Society, the Graphic Design and Molecular Modeling Group, and the CNRS Working Group on Symbolic Systemic Biology (GDRs IM and BIM).

Bioinformatic tools to analyze nonribosomal peptides and other metabolites

Maude PUPIN

Université de Lille, 59650, Villeneuve d'Ascq cedex, France

Invited speaker: maude.pupin@univ-lille.fr

Non-ribosomal peptides are natural peptides synthesized by specialized enzymes, the synthetases, encoded by large gene clusters. This synthetic pathway produces a great diversity of compounds because it selects and incorporates not only several hundred different amino acids, but also fatty acids or monosaccharides. These basic building blocks are called monomers. Moreover, bonds of different kinds are made between monomers, leading to various structures that may contain cyles and/or branches.

I will present the Norine software platform dedicated to these molecules. It proposes a specific notation to describe the monomer structure: a graph in which the nodes are the monomers and the edges are the chemical bonds between them. This notation facilitates the association between a metabolite and its biosynthetic gene cluster. It can be used for any metabolite synthesized by the assembly of monomers.

Norine offers tools to analyze and visualize secondary metabolites based on this notation. s2m and rBAN deduce a monomeric structure from a chemical structure. Kendrick Formula Predictor and NRPro help analyze mass spectrometry results. And Norine is a database dedicated to non-ribosomal peptides (see https://bioinfo.cristal.univ-lille.fr/NRP/).

Multi-species predictive metabolomics via GLM approach unveils a generic metabolic toolbox for plant response to an extreme abiotic gradient in the Atacama Desert

Thomas DUSSARRAT[1,2], Sylvain PRIGENT[2], Claudio LATORRE[4], Stéphane BERNILLON[2], Amélie FLANDIN[2], Francisca DIAZ[1], Cédric CASSAN[2], Pierre VAN DELFT[3], Daniel JACOB[2], Jérôme JOUBES[3], Yves GIBON[2], Dominique ROLIN[2], Rodrigo A. GUTIÉRREZ[1] and Pierre PÉTRIACQ[2]

[1] Departamento de Genética Molecular y Microbiología, Pontificia Universidad Católica de Chile, FONDAP Center for Genome Regulation and Millenium Institute for Integrative Biology (iBio), Av Libertador Bernado O'Higgins 340, Santiago, Chile

[2] Univ. Bordeaux, INRAE, UMR 1332 BFP, 71 av E. Bourlaux, 33140, Villenave d'Ornon, France

[3] Univ. Bordeaux, CNRS, UMR 5200 Laboratoire de Biogenèse Membranaire, 71 av E. Bourlaux, 33140, Villenave d'Ornon, France

[4] Departamento de Ecología, Pontificia Universidad Católica de Chile, Av Libertador Bernado O'Higgins 340, Santiago, Chile

Invited speaker: thomas.dussarratbiologie@gmail.com

Performances of the best ideotypes are threatened by the increased aridity worldwide. Developing new strategies to improve our understanding of abiotic plant tolerance is urgent for more sustainable agriculture. Wild plant species that inhabit extreme lands such as deserts and high mountains represent a unique resource of adaptive molecular mechanisms. Yet, plant adaptation to hostile biomes is mainly considered species-specific [1] . However, the contribution of generic mechanisms remains unexplored. Here, we propose a comprehensive approach to investigate the role of metabolic processes in the adaptation of multiple selected species from the Atacama Desert, the driest non-polar environment on earth [2] . First, we compared gene expansion and expression patterns between 32 selected Atacama species thriving on an elevation gradient from 2500 to 4500m and 32 closest sequenced species. Reaction enrichment analyses via Pathway Tools identified genetic evolutions governing convergent biochemical pathways. Subsequently, the biochemical diversity of Atacama species was accessed using multi-platform metabolomics. Metabolome and plant environment were linked through generalised linear models to study the role of shared metabolic mechanisms in plant resilience. This predictive metabolomics approach unveiled a generic metabolic toolbox predicting plant environment, independently of plant species and year [3]. The predictive metabolites were detected in agronomic and ornamental species. Overall, while this study enhanced our understanding of the place of generic mechanisms in adaptation to extreme climates, our multi-species approach combined with machine learning offered promising perspectives in agronomy and ecology.

References

1. Scossa, F. & Fernie, A. R. The evolution of metabolism: How to test evolutionary hypotheses at the genomic level. Computational and Structural Biotechnology Journal 18, 482–500 (2020)
2. Eshel, G. et al. Plant ecological genomics at the limits of life in the Atacama Desert. Proc Natl Acad Sci USA 118, e2101177118 (2021)
3. Dussarrat, T. et al. Predictive metabolomics of multiple Atacama plant species unveils a core set of generic metabolites for extreme climate resilience. New Phytologist nph.18095 (2022) doi:10.1111/nph.18095

Bayesian matrix factorization and deep learning for drug-target activity prediction

Yves MOREAU

KU Leuven, ESAT-STADIUS, Kasteelpark Arenberg 10, 3001, Heverlee, Belgium

Invited speaker: yves.moreau@esat.kuleuven.be

Learning latent representations via matrix completion or deep learning provides an attractive framework to handle sparsely observed data, also called “scarce” data. A typical setting for scarce data is the prediction of biological activity of chemical compounds against drug targets, where only 0.1% to 1% of all compound-target pairs are measured. Matrix factorization searches for latent representations of compounds and targets that allow an optimal reconstruction of the observed measurements. These methods can be further combined with linear regression models to create multitask prediction models. In our case, fingerprints of chemical compounds are used as “side information” to predict target activity. Such representations can also be leveraged to link drug targets and pathway activation. Similar results can be achieved via deep learning. Next to classical chemical fingerprints, high-content imaging can also be used to predict drug-target activity with improved performance for scaffold hopping. Our methods are available as two open source Python/C++ libraries - Macau & SparseChem (https://github.com/jaak-s/macau/tree/master/python/macau; https://github.com/melloddy/SparseChem).

Symposium 4, Amphi D

Simulation of biological sequences and evolutionary processes

Simulation of biological sequences

Organizers: Guy PERRIÈRE[1], Guillaume BESLON[2,3], Bastien BOUSSAU[1], Jonathan ROUZAUD-CORNABAS[2,3], Vincent DAUBIN[1] and Eric TANNIER[1,3]

[1] Laboratoire de Biométrie et Biologie Évolutive – UMR CNRS 5558, Université Claude Bernard – Lyon 1, Villeurbanne, France

[2] Laboratoire d’Informatique en Image et Systèmes d’Information – UMR CNRS 5205, INSA de Lyon, Villeurbanne, France

[3] Projet INRIA BEAGLE – Artificial Evolution and Computational Biology, INRIA Antenne Lyon la Doua, Villeurbanne, France

This mini-symposium is supported by ANR Evolution.

Cophylogeny reconstruction in host-parasites systems via Approximate Bayesian Computation and allowing for multiple associations

Catherine MATIAS

Laboratoire de Probabilités, Statistique et Modélisation, UMR CNRS
8001, Sorbonne Université, Université de Paris, 75252, Paris, France

Invited speaker: catherine.mathias@math.cnrs.fr

Nowadays, the most used method in studies of the coevolution of hosts and symbionts is phylogenetic tree reconciliation. Most algorithms for host–parasite reconciliation use an event-based model, where the events include in general (a subset of) cospeciation, duplication, loss, and host switch. All known parsimonious event-based methods then assign a cost to each type of event in order to find a reconstruction of minimum cost. A crucial issue is that from a biological point of view, reasonable cost values for an event-based parsimonious reconciliation are not easily chosen, and this choice strongly influences the reconciliation obtained. Thus, different approaches have been developed to infer such cost values for a given pair of host and symbiont trees.

We developed an algorithm, called Coala, for estimating the frequency of the events based on an approximate Bayesian computation (ABC) approach. The benefits of this method are 2-fold: i) it provides more confidence in the set of costs to be used in a reconciliation; and ii) it allows estimation of the frequency of the events in cases where the data set consists of trees with a large number of taxa.

However, a major limitation of all these approaches (including Coala) is their inability to model the 'invasion' of different host species by a same symbiont species (referred to as a spread event), which is thought to happen in symbiotic relations.

To mention one example, the same species of insects may pollinate different species of plants. This results in multiple associations observed between the symbionts and their hosts (meaning that a symbiont is no longer specific to a host), that are not compatible with the current methods of coevolution. After presenting the context and the tool Coala, I will discuss an extension called Acoala, which for a given pair of host and symbiont trees, estimates the probabilities of the cophylogeny events, in presence of spread events, still relying on an ABC approach.

The algorithm that we propose, by including spread events, enables the multiple associations to be taken into account in a more accurate way, inducing more confidence in the estimated sets of costs and thus in the reconciliation of a given pair of host and symbiont trees. Its rooting in the tool Coala allows it to estimate the probabilities of the events even in the case of large datasets. We will illustrate the performance of the method on real host-symbionts associations datasets.

Deep convolutional and conditional neural networks for large-scale genomic data generation

Burak YELMEN[1,2]

[1] INRIA, Laboratoire Interdisciplinaire des Sciences du Numérique, 91400, Orsay, France
[2] Institute of Genomics, University of Tartu, 51010, Tartu, Estonia; Institute of Molecular and Cell Biology, University of Tartu, 51010, Tartu, Estonia

Invited speaker: Burak.Yelmen@lisn.fr

Applications of generative models for genomic data have gained significant momentum in the past few years, with scopes ranging from data characterization to generation of simulated genomes. Genomic data is structured and a variety of processes such as recombination, mutations, natural selection and past demographic events contribute to the complexity of genomic data structure. In our previous study, we demonstrated that generative adversarial networks (GANs) and restricted Boltzmann machines (RBMs) can be used to create novel high-quality artificial genomes which can preserve the complex characteristics of real genomes such as population structure, linkage disequilibrium and selection signals. However, a major drawback of these models is scalability, as the large feature space of genome-wide data increases computational complexity vastly.

To address this issue, we explored different convolutional GAN and variational autoencoder (VAE) architectures along with a new conditional RBM framework for generating artificial genomes with high SNP coverage. All these networks implicitly learn the varying landscape of recombination in order to capture complex correlation patterns along the genome and generate a wide diversity of plausible haplotypes. We performed comparative analyses to assess both the quality of these generated haplotypes and the amount of privacy leakage from the training data. As the importance of genetic privacy becomes more prevalent, the need for effective privacy measures for genomic data increases. In the near future, large-scale artificial genome databases can be assembled to provide easily accessible surrogates of real databases, allowing researchers to perform studies with diverse genomic data within a safe ethical framework.

Simulate genomic sequences evolution along complex phylogenetic trees with Aevol

Guillaume BESLON[1,2]

[1] Laboratoire d'Informatique en Image et Systèmes d'Information – UMR CNRS 5205, INSA de Lyon, Villeurbanne, France

[2] Projet INRIA BEAGLE – Artificial Evolution and Computational Biology, INRIA Antenne Lyon la Doua, Villeurbanne, France

Invited speaker: guillaume.beslon@inria.fr

Simulating the evolution of a biological sequence requires the integration of mutation, selection and genetic drift processes. However, the coupling between these three processes is based on the possibility of attributing a reproductive capacity to any sequence likely to appear – by mutation – within the population. Several approaches have thus been proposed, including "artificial chemistry" which define an abstract algorithmic process for decoding the genome and simulate evolution in the virtual biochemical world generated. However, these approaches generate sequences that have little in common with real biological sequences: thus, Avida, one of the best-known software in this field, simulates sequences composed of assembler pseudo-instructions whose biological interpretation raises questions, to say the least. To overcome this difficulty, one option would be to adopt an artificial chemistry model directly inspired by the "genotype-to-phenotype map" and thus allowing to observe experimentally (although in silico) the evolution of realistic sequences. The central idea here is that the effects of mutations are, in first approximation, dictated by the structure of genomes and therefore by their decoding process.

I will present Aevol, a simulator based on this idea, as well as some recent results obtained with this model. In particular, I will show how this model can simulate the evolution of populations of genomic sequences along more or less complex phylogenetic trees and how, within the framework of the Evoluthon project, these trees could then be reconstructed using classical (and independent of the model) bioinformatics tools. Without formally constituting a validation of the model, this successful coupling between simulation and phylogenomic analysis suggests that, in Aevol, the sequences have evolved by a process presenting a certain – but unquantifiable – degree of realism.

Simulating SARS-CoV-2 Sequences: phylogenies and metagenomes

William BOULTON

School of Computing Sciences, University of East Anglia, Norwich,
United Kingdom

Invited speaker: W.Boulton@uea.ac.uk

For bioinformaticians, an inspiring challenge arising from the COVID-19 pandemic has been the incredible wealth of data generated by sequencing SARS-CoV-2 genomes from clinical cases, surely the largest epidemiological dataset of its kind [1,2]. In many countries, these data were supplemented by wastewater surveillance, which acted as a bellwether for clinical case numbers and promises to allow for strain-resolved metagenomics to estimate proportions of variants of concern [3].

Many bioinformaticians, including our group, immediately shifted focus to analysing this deluge of new viral sequences, and several groups developed novel algorithms to analyse SARS-CoV-2 phylogenies and metagenomes, such as UShER, iVar and Freyja [4–6]. However, existing tools were not well-equipped to simulate these data, a necessary step when benchmarking new methods. We developed two tools to alleviate these problems: phastSim [7], a fast genome sequence simulator for pandemic scale phylogenies, and SWAMPy [8], a simulator for SARS-CoV-2 metagenomes generated by amplicon sequencing of wastewater samples. This talk will provide a summary of the methods and novel features of these two simulation tools.

References

1. Chen Z., Azman A.S., Chen X. et al. (2022) Global landscape of SARS-CoV-2 genomic surveillance and data sharing. Nat. Genet. 54:499–507. doi:10.1038/s41588-022-01033-y
2. Hodcroft E.B., De Maio N., Lanfear R. et al. (2021) Want to track pandemic variants faster? Fix the bioinformatics bottleneck Nature 591:30–33. doi:10.1038/d41586-021-00525-x
3. Baaijens J.A., Zulli A., Ott I.M. et al. (2021) Variant abundance estimation for SARS-CoV-2 in wastewater using RNA-Seq quantification. medRxiv 2021.08.31.21262938. doi:10.1101/2021.08.31.21262938
4. Turakhia Y., Thornlow B., Hinrichs A.S. et al. (2021) Ultrafast Sample placement on Existing tRees (UShER) enables real-time phylogenetics for the SARS-CoV-2 pandemic. Nat. Genet. 53:809–816. doi:10.1038/s41588-021-00862-7
5. Grubaugh N.D., Gangavarapu K., Quick J. et al. (2019) An amplicon-based sequencing framework for accurately measuring intrahost virus diversity using PrimalSeq and iVar. Genome Biol. 20:8. doi:10.1186/s13059-018-1618-7
6. Karthikeyan S, Levy J.I., De Hoff P. et al. (2021) Wastewater sequencing uncovers early, cryptic SARS-CoV-2 variant transmission. medRxiv 2021.12.21.21268143. doi:10.1101/2021.12.21.2126814
7. De Maio N., Boulton W., Weilguny L. et al. (2022) phastSim: Efficient simulation of sequence evolution for pandemic-scale datasets. PLoS Comput. Biol. 18:e1010056. doi:10.1371/journal.pcbi.1010056
8. https://github.com/goldman-gp-ebi/sars-cov-2-metagenomic-simulator

Symposium 5, Amphi E

Structural characterization of macro-assemblies using integrative methods

Structural characterization of macro-assemblies using integrative methods

Organizers: Isaure CHAUVOT DE BEAUCHENE[1], Jessica ANDREANI[2], Anne LOPES[2] and Matthieu MONTES[4]

[1] Université de Lorraine, CNRS, Inria, LORIA, F-54000, Nancy, France

[2] Institute for Integrative Biology of the Cell (I2BC), Université Paris-Saclay, 91198, Gif-sur-Yvette, France

[3] Laboratoire GBCM, EA 7528, CNAM, Hésam Université, 75003, Paris, France

This mini-symposium is supported by:

Contradictory tips for integrative modeling of Cryo-Electron Microscopy Maps

Pablo CHACON

Department of Biological Physical Chemistry, Rocasolano Physical
Chemistry Institute C.S.I.C., E-28006 Madrid, Spain

Invited speaker: pablo@chaconlab.org

We have witnessed considerable advancements over the past few years in both prediction structures using machine learning and structure determination using Cryo-Electron Microscopy (Cryo-EM). Commenting on my recent experience as a modeler, I briefly introduce the impact of such advances in the integrative modeling of Cryo-EM electron density maps. In this context I discuss two systems 1) tyrosine hydroxylase inhibition by dopamine and reactivation by Ser40 phosphorylation and 2) and continuous flexibility analysis of SARS-CoV-2 Spike prefusion structures.

Integrative structural biology through the lens of structural bioinformatics

Guillaume BOUVIER

Institut Pasteur, Unité de Bioinformatique Structurale, CNRS UMR 3528,
Département de Biologie Structurale et Chimie, F-75015, Paris, France

Invited speaker: guillaume.bouvier@pasteur.fr

Structural bioinformatics is a field of research at the interface of multiple areas. To create structural models of biological macromolecules, ones often need to integrate heterogeneous data coming from multiple experiments; e.g., Nuclear Magnetic Resonance (NMR), X-ray crystallography, Small Angle X-ray Scattering (SAXS), Cryo-Electron Microscopy (Cryo-EM), Chemical cross-linking in combination with mass spectrometry (XL-MS)… The combination of such heterogeneous data into a structural model requiresthe development of modeling methods to process them and extract meaningful information from them.

Furthermore, the recent breakthrough in deep-learning approaches applied in the field of structural biology has remodeled the way of thinking about problem-solving strategies. I will expose the new methodological developments we made for the structural characterization of large biomolecular assemblies. I will start from the traditional approaches then turn out into the recent development we made in the field of machine-learning and especially deep-learning.

Crystallographer Adventures in Wonderland

Pierre LEGRAND

Synchrotron SOLEIL, 91192, Gif-sur-Yvette, France

Invited speaker: pierre.legrand@synchrotron-soleil.fr

Last summer, as AlphaFold2 hurried by, some crystallographers burning with curiosity ran after it across a new field and, eventually, jumped into a large AI-Hole. Down, down, down, they felt as if the fall would never come to an end. We will share the journey of one of them in this wonderful new deep field and hear what he discovered. Although he half believed the wonders he saw, in the after-time he pictured to himself how this could help him fertilize his own experimental garden.

Simulating and understanding the Extracellular Matrix. From isolated elements to reconstructed pictures: key role of the in silico approach.

Stéphanie BAUD

CNRS UMR 7369, Matrice Extracellulaire et Dynamique Cellulaire, Reims, 51100, France;
P3M, Multi-scale Molecular Modeling Plateform, Université de Reims Champagne Ardenne, 51100, Reims, France

Invited speaker: stephanie.baud@univ-reims.fr

The extracellular matrix (ECM) is a three-dimensional network of macromolecules that is the architectural support for cells and allows tissue cohesion. This dynamic structure regulates many biological functions such as adhesion, migration, proliferation, differentiation and cell survival. Four main families of macromolecules constitute this interstitial medium: collagens, structural glycoproteins, proteoglycans and elastins.

From the modeling point of view, the ECM remains a challenging object: indeed its study can be approached according to different aspects and scales (quantum, atomic, molecular or even mesoscopic) which nevertheless remain complementary. According to the framework of investigation, the diversity of size, function and nature of the ECM molecules imposes to choose a given scale of description. Indeed, there are still few methods available in physics that can describe/model a complex system using a unified approach combining physico-chemical properties linked to different descriptive scales.

Using simulation and modeling from the atomistic scale up to mesoscopic scale, our team is aiming at understanding and deciphering the structural and dynamic behavior of the ECM. In particular, the recent research projects we have been developing focus on a better description of the ECM, the interactions with its microenvironment as well as the impact of its modifications (natural such as the phenomenon of aging, or pathological such as diabetic or cancerous contexts). In particular, the detailed description of the interactions between the various constituent elements of the ECM and the finalization of a modeling tool adapted to the mesoscopic scale have been investigated.

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