JOBIM 2025

08>11 JUIL

Bordeaux

Proceedings

Preface

Dear Participants and Colleagues,

We are very pleased to present the proceedings of the 2025 edition of JOBIM, hosted this year in Bordeaux, which brought together 500 participants on-site and more than 90 joining online.

This edition stands out in particular for the fruitful collaboration between the Program Committee and PCI Genomics and PCI Mathematical and Computational Biology, resulting in 9 submissions through the PCI track. All PCI submissions were accepted for presentation at JOBIM and have entered the PCI peer-review process, an encouraging sign of the growing synergy between open peer-review platforms and scientific conferences.

Thanks to the dedication of the Program Committee, we were able to finalize a rich and balanced scientific program. In total, we received 61 long submissions, leading to 22 accepted Proceedings articles out of 34 submissions, 9 PCI submissions (all accepted for presentation), 5 long platform papers accepted out of 6, 9 Highlights accepted out of 12, helping ensure thematic diversity. This results in 45 oral presentations, scheduled across 15 parallel sessions. In addition, 151 posters were accepted out of 176 submissions.

We are pleased to highlight the nine mini-symposia proposals received this year, of which four were selected by the Program Committee. These four themes will enrich Thursday afternoon with an exciting and diverse program, covering:

- *Comment concilier nos activités en bioinformatique avec les limites planétaires ?*
- *The Genomics of Biodiversity*,
- *AI in Healthcare: From Fundamentals to the Clinic*, and
- *Methods for Interfacing with Graphs of Genomic Sequences: Novel Pangenome Paradigms*.

We warmly thank all reviewers, authors, and contributors who made this scientific program possible. We hope these proceedings reflect the diversity, innovation, and collaborative spirit of the JOBIM 2025 community.

In addition, we are particularly honored to welcome six distinguished keynote speakers whose contributions have greatly enriched this edition of JOBIM: Anamaria Necsulea (CNRS, Lyon), Björn Grüning (University of Freiburg), Simona Cocco (CNRS, ENS Paris), Emma Schymanski (University of Luxembourg), Eric Rivals (CNRS, Montpellier) and Jean Monlong (INSERM, Toulouse).

We would also like to express our sincere thanks to our institutional partners: SFBI, IFB and GdR BIMMM. Their support and commitment have been essential to the success of this event.

Finally, we thank everyone who contributed to making JOBIM 2025 a welcoming and vibrant space for the exchange of ideas in computational biology.

We hope you will enjoy these proceedings, and we look forward to your continued engagement in the future editions of JOBIM.

Sylvain Prigent, Patricia Thébault and Raluca Uricaru

Presidents of the Organizing Committee

Romain Bourqui	Université de Bordeaux	Bordeaux
Clémence Frioux	INRIA	Bordeaux
Simon Labarthe	INRAE /INRIA	Bordeaux

Members of the Organizing Committee

Yanis Asloudj	Université de Bordeaux	Bordeaux
David Auber	Université de Bordeaux	Bordeaux
David Benaben	INRAE	Villenave d'Ornon
Guillaume Blin	Université de Bordeaux	Bordeaux
Laetitia Bourgeade	CHU	Bordeaux
Victoria Bourgeais	Université de Bordeaux	Bordeaux
Romain Bourqui	Université de Bordeaux	Bordeaux
Catherine Cattaert-Megrat	INRIA	Bordeaux
Emilie Chancerel	INRAE	
Antonin Colajanni	Université de Bordeaux	Bordeaux
Carole Couture	INRAE	Villenave d'Ornon
Elodie Darbo	Université de Bordeaux	Bordeaux
Cyril Dourthe	Université de Bordeaux	Bordeaux
Jean-Marc Frigerio	INRAE	Villenave d'Ornon
Clémence Frioux	INRIA	Bordeaux
Virginie Garcia	INRAE	Villenave d'Ornon
Romain Giot	Université de Bordeaux	Bordeaux
Loann Giovannelli	Université de Bordeaux	Bordeaux
Jean-Christophe Helbling	INRAE	Pessac
Simon Labarthe	INRAE	Cestas - Pierroton
Malo Le Boulch	INRAE	Villenave d'Ornon
Pascal Martin	INRAE	Villenave d'Ornon
Fleur Mougin	Université de Bordeaux	Bordeaux
Loïc Paulevé	CNRS	Bordeaux
Nadia Ponts	INRAE	Villenave d'Ornon
Sylvain Prigent	INRAE	Villenave d'Ornon
Franck Salin	INRAE	Cestas - Pierroton
Joris Sansen	INP ENSTBB	Bordeaux
Patricia Thébault	Université de Bordeaux	Bordeaux
Nicolas Tourasse	CNRS	Bordeaux
Joseph Tran	INRAE	Villenave d'Ornon
Raluca Uricaru	Université de Bordeaux	Bordeaux

Presidents of the Program Committee

Sylvain Prigent	INRAE	Villenave d'Ornon
Patricia Thébault	Université de Bordeaux	Bordeaux
Raluca Uricaru	Université de Bordeaux	Bordeaux

Members of the Program Committee

Sophie	**Abby**		CNRS	Grenoble
Julie	**Aubert**		INRAE	Paris
Benoît	**Ballester**		INSERM	Marseille
Anais	**Bardet**		CNRS	Strasbourg
Millena	**Barros-Santos**		INRAE	Avignon
Anaïs	**Baudot**		CNRS	Marseille
Séverine	**Bérard**		Université	Montpellier
Camille	**Berthelot**		Pasteur	Paris
Yuna	**Blum**		CNRS	Rennes
Jérémie	**Bourdon**		Université	Nantes
Anne-Claude	**Camproux**		Université	Paris
Frédéric	**Cazals**		INRIA	Sophia-Antipolis
Bastien	**Cazaux**		Université	Lille
Isaure	**Chauvot de Beauchêne**		INRIA	Nancy
Hélène	**Chiapello**	IFB	INRAE	Paris
Rayan	**Chikhi**		Pasteur	Paris
Erwan	**Corre**	IFB	CNRS	Roscoff
Olivier	**Dameron**		Université	Rennes
Elodie	**Darbo**	SFBI	INSERM	Bordeaux
Sarah	**Djebali**		INSERM	Toulouse
Patrick	**Durand**		IFREMER	Plouzané
Damien	**Eveillard**		Université	Nantes
Anna-Sophie	**Fiston-Lavier**	SFBI	Université	Montpellier
Clovis	**Galiez**		Université	Grenoble
Christine	**Gaspin**		INRAE	Toulouse
Franck	**Giacomoni**		INRAE	Clermont
Emmanuel	**Giudice**		Université	Rennes
Fabien	**Jourdan**		INRAE	Toulouse
Slim	**Karkar**		Université	Bordeaux
Vincent	**Lacroix**		Université	Lyon
Sandrine	**Lagarrigue**		INRAE	Rennes

Elodie	**Laine**		Université	Paris
Yann	**Le Cunff**		Université	Rennes
Charles	**Lecellier**		CNRS	Montpellier
Claire	**Lemaitre**	GDR BIMMM	INRIA	Rennes
Emmanuelle	**Lerat**		CNRS	Lyon
Camille	**Marchet**		CNRS	Lille
Mahendra	**Mariadassou**	GDR BIMMM	INRAE	Paris
Raphael	**Mourad**		Université	Toulouse
Anna	**Niarakis**		Université	Toulouse
Loïc	**Paulevé**		CNRS	Bordeaux
Eric	**Pelletier**		CEA	Paris
Sabine	**Peres**		Université	Lyon
Yann	**Ponty**		CNRS	Paris
Delphine	**Pottier**		CNRS	Marseille
Nathalie	**Poupin**		INRAE	Toulouse
Céline	**Scornavacca**		CNRS	Montpellier
Nathalie	**Vialaneix**		INRAE	Toulouse
Matthias	**Zytnicki**		INRAE	Toulouse

List of Presentations

Keynote speakers

Integrating structural variants in genomic studies of rare and complex diseases with long-read sequencing and pangenomes

Jean MONLONG[1]

1 Institut de Recherche en Santé Digestive, Université de Toulouse, INSERM, INRA, ENVT, UPS, Toulouse, France

Corresponding Author: jean.monlong@inserm.fr

Keywords

structural variants, sequencing, pangenome, rare disease

Abstract

Variant affecting more than 50 nucleotides, or structural variants, can have important functional impacts. They are unfortunately understudied because of technical challenges hindering their detection. I will present two approaches to integrate those variants in genomic studies. The first uses pangenomes as augmented reference genome containing common variants (including structural variants). With this more complete reference, like the one produced by the Human Pangenome Reference Consortium, short sequencing reads are better analyzed, resulting in a larger number of structural variants that can finally be genotyped accurately and could be considered in large association studies. The second approach uses cost-efficient long read sequencing technology, such as Oxford Nanopore, to infer phased variants at unprecedented resolution. This protocol is currently being tested to help diagnose rare disease patients in diagnostic impasse.

On n'a pas de terres rares, mais on a des idées.

- Partitionnement en classe d'équivalence pour le test d'algorithmes du texte.
- Vers une approche scientifique pour choisir au mieux les instances de tests.

Eric Rivals[1]

[1] LIRMM, CNRS, Univ Montpellier, FRANCE

Corresponding Author: eric.rivals@lirmm.fr

Abstract

La consommation de ressources, telles que l'électricité ou les matériaux de fabrication des ordinateurs, s'accroît fortement avec le développement de l'économie et des activités humaines numériques. La question se pose de limiter cette consommation sans pour autant réduire des activités qui peuvent s'avérer socialement utiles ou nécessaires.

En bioinformatique, nous développons beaucoup de logiciels, et parfois de nombreux logiciels pour la même tâche, la même question computationnelle. Prenez par exemple le cas de l'assemblage de génome ou celui de la localisation des lectures de séquençage (ou "read mapping" en anglais--footnote{Plusieurs dizaines d'outils de mapping ont été développés et maintenus.}).

En matière de développement logiciel, il est recommandé de développer des tests les plus complets possibles afin de s'assurer de la validité d'un logiciel. Dans le cycle de développement, nous itérons, fréquemment ou automatiquement, l'exécution des tests afin de vérifier la correction du logiciel ou d'en évaluer la rapidité. Dès lors, on peut se questionner scientifiquement sur l'utilité ou la redondance de certaines instances de tests.

Considérons le cas d'un programme de recherche d'un mot dans un texte, par exemple votre génome préféré (ou bien dans la séquence de De Bruijn d'ordre k). Pour tester le logiciel dans toutes les situations, on peut lancer des tests sur tous les mots de longueur k, pour k égale 2, puis 3, puis 4, ..., jusqu'à par ex. disons 31 (footnote{La valeur k=31 est commune pour certaines analyses sur séquence d'ADN.}). Mais clairement le nombre d'instances augmente exponentiellement avec la longueur k, et devient vite rédhibitoire. Du point de vue de la science informatique, nous pouvons reformuler notre question initiale ainsi:

Peut-on identifier des instances de test redondantes ? Peut-on générer seulement des instances de test utiles ?

J'aborderai ces questions durant cet exposé en les illustrant avec deux algorithmes complexes de traitement des séquences et en montrant l'impact du choix des instances. Ce principe d'organisation

des tests se nomme Partitionnement par Classes d'Équivalence. L'étude de ces classes d'équivalence pour un algorithme donné peut s'avérer complexe.

Étant donnée la fréquence d'exécution des tests durant le développement logiciel, ce type d'approche peut aider à diminuer l'impact écologique de nos développements. Outre l'avantage en termes d'utilisation de ressources, l'approche par Partitionnement peut aussi nous informer sur le temps moyen d'exécution du programme sur une classe d'instances d'une taille donnée. Cette réflexion générique ouvre des pistes de recherches pour de nombreux algorithmes, pistes à même de favoriser des interactions avec d'autres domaines de la recherche en informatique.

Abstract

Selection of test instances for string algorithms

The consumption of resources, such as electricity and materials used in the manufacture of computers, is increasing sharply with the development of the digital economy and human activities. The question is how to limit this consumption without reducing activities that may be socially useful or necessary.

In bioinformatics, we develop a lot of software, and sometimes many software for the same task, the same computational question. Think, for instance, about genome assembly or read mapping.

When it comes to software development, it's advisable to develop the most comprehensive series of tests to ensure the validity of a piece of software. In the development cycle, we iterate, frequently or automatically, the execution of tests in order to verify the software's correctness at each step. Tests may also be run to evaluate computational speed. This raises scientific questions about the usefulness or redundancy of certain test instances.

Let's take the case of a program that searches for a word in a text, such as your favorite genome. To test the program in all situations, we can run tests on all words of length k, for k equals 2, then 3, then 4, ..., up to, say 31 (footnote{The value k=31 is common for some DNA sequence analyses.}). But clearly the number of instances increases exponentially with length k, and quickly becomes prohibitive. From the point of view of computer science, we can rephrase our initial question as follows:

Can we identify redundant test instances? Can we generate only useful test instances?

I will address these questions during this talk, illustrating them with two complex sequence processing algorithms and showing the impact of instance selection.

This principle of test organization is called Equivalence Class Partitioning ([[https://en.wikipedia.org/wiki/Equivalence_partitioning][ECP]]) in the domain of software development. The study of these equivalence classes for a given algorithm can be complex.

Given the frequency with which tests are run during software development, and the number of software in bioinformatics, this type of approach can help reducing the ecological impact of our developments.

In addition to the advantage in terms of resource usage, the partitioning approach can also inform us about the average program execution time on a class of instances of a given size. This generic approach points to avenues of research for many core algorithms, avenues that may foster interactions with other areas of computer science.

LocalWords: reads Bruijn

Deciphering the genomic basis of convergent phenotypic evolution

Anamaria Necsulea[1]

1 Laboratoire de Biométrie et Biologie Evolutive, Centre National de la Recherche Scientifique, Université Claude Bernard – Lyon 1

Corresponding Author: anamaria.necsulea@univ-lyon1.fr

Keywords

convergent phenotypic evolution, evolutionary genomics, gene expression evolution, regulatory evolution.

Abstract

Biology has recently undergone a fundamental transformation, powered by the development of sensitive molecular techniques combined with high-throughput sequencing. Major technical breakthroughs include the ability to sequence complete genomes, to finely quantify gene activity and to study its control mechanisms. Applying these techniques across species gives us a unique opportunity to study the evolution of genomic functions, and thus to better understand the mechanisms that underlie phenotypic evolution. Functional evolutionary genomics studies can bring insights into the selective pressures and molecular mechanisms that drive the emergence (or loss) of biological functions and of phenotypes.

Here, we illustrate how comparative genomics approaches can bring insights into the genomic basis of convergent phenotypic evolution, in birds. The avian clade displays a spectacular diversity of phenotypes. Numerous instances of convergent phenotypic evolution are known in birds, such as the convergent loss of flight [1] or the parallel gain of vocal learning [2]. In this presentation, we will focus on one peculiar case of convergent morphological evolution : the loss of the intromittent male phallus [3]. Although an intromittent phallus was likely present in the ancestor of all amniotes, this organ was reduced or entirely lost in multiple avian lineages, including the major Neoaves clade and the Phasianidae family [3]. The evolutionary processes that led to phallus reduction or loss are still unclear, as are the genomic consequences of this major phenotypic change. Taking advantage of the availability of hundreds of avian genomic sequences, we have performed large-scale evolutionary analyses of protein-coding gene sequences and of non-coding regulatory elements, searching for genomic changes that occur in parallel with phenotypic changes. We found that hundreds of protein-coding

genes and non-coding regulatory elements underwent an acceleration of their rate of evolution following this major phenotypic change. We also identify numerous gene expression differences between bird species that have retained the intromittent phallus and species that have lost this organ. While we cannot claim that these changes in expression patterns and regulatory programs are causal to the loss of the phallus, our findings illustrate the genome-wide consequences of this major phenotypic change.

References

1. Pan S, Lin Y, Liu Q, Duan J, Lin Z, Wang Y, et al. Convergent genomic signatures of flight loss in birds suggest a switch of main fuel. Nat Commun. 21 juin 2019;10(1):2756.

2. Pfenning AR, Hara E, Whitney O, Rivas MV, Wang R, Roulhac PL, et al. Convergent transcriptional specializations in the brains of humans and song-learning birds. Science. 12 déc 2014;346(6215):1256846.

3. Herrera AM, Shuster SG, Perriton CL, Cohn MJ. Developmental basis of phallus reduction during bird evolution. Curr Biol CB. 17 juin 2013;23(12):1065 74.

Open Science Data Processing and Integration Workflows in Metabolomics and Exposomics

Emma Schymanski[1]

1 Luxembourg Centre for Systems Biomedicine, University of Luxembourg, 6 avenue du Swing, 4367, Belvaux, Luxembourg.

Corresponding Author: emma.schymanski@uni.lu

Keywords

Metabolomics, Exposomics, Open Science, Cheminformatics, Non-target Screening.

Abstract

Exposomics researchers need to identify relevant chemicals covering the entirety of potential exposures over entire lifetimes. With over 100 million chemicals in the largest open chemical databases, coupled with broadly acknowledged knowledge gaps, researchers are faced with too much yet not enough information at the same time. Improvements in analytical technologies and computational mass spectrometry workflows coupled with the rapid growth in databases and increasing demand for high throughput "big data" services from the research community present significant challenges for both data hosts and workflow developers. This talk will showcase FAIR and Open Science developments in the Environmental Cheminformatics group, including the NORMAN Suspect List Exchange (NORMAN-SLE), MassBank, MetFrag, PubChemLite for Exposomics [1], patRoon, ShinyTPs and the Chemical Stripes. Beyond the software developments, it will showcase how these are applied in our active research projects in our data processing and integration workflows to tackle challenges in non-target exposomics studies [2,3]. The case studies will show how enhancing the FAIRness (Findability, Accessibility, Interoperability and Reusability) of open resources can mutually enhance several resources for whole community benefit. Many thanks to all group members, collaborators and colleagues who have been a part of these efforts!

References

1. Schymanski EL et al. Empowering large chemical knowledge bases for exposomics: PubChemLite meets MetFrag, Journal of Cheminformatics, 2021; 13:19.

2. Talavera Andújar B et al. (2024) Can Small Molecules Provide Clues on Disease Progression in Cerebrospinal Fluid from Mild Cognitive Impairment and Alzheimer's Disease Patients? Environ. Sci. Technol, 2024 ; 58(9) :4181-41923.

3. Talavera Andújar B et al. Exploring environmental modifiers of LRRK2-associated Parkinson's disease penetrance: An exposomics and metagenomics pilot study on household dust. Environment International, 2024; 194:109151 DOI: 10.1016/j.envint.2024.109151

Generative models learned on sequence data to forecast SARS-CoV-2 viral evolution and antibody resilience.

Simona COCCO

LPENS, 24 rue Lhomond, 75005, Paris, France

simona.cocco@phys.ens.fr

Keywords

Restricted Boltzmann Machines, SARS-CoV-2, antibody escape, deep mutational scanning, antibody resilience, viral adaptation.

Abstract

In this talk I will introduce the Restricted Boltzmann Machines (RBM): a simple Machine learning model with a bipartite graph architecture, learned only on sequence data. I will focus on the applications of RBM on the predictions of SARS-CoV-2 evolution. By integrating pre-pandemic evolutionary constraints gathered from SARS-CoV-2 far homologous sequences, with large-scale Deep mutational Scans (DMS) data we model how viral fitness, ACE2 binding, and immune escape pressures jointly sculpt the mutational landscape. In contrast to structure-based models that are restricted in scalability, our sequence-based energy framework enables broad exploration of evolutionary trajectories while remaining valuable for experimental validation. Experimental validation of model predictions includes the test of 22 synthetic RBD variants with up to 21 mutations from the wild-type. Half of these variants maintained expression and ACE2 binding, and some successfully escaped most of the 9 antibodies tested.

Bibliography

[1] Minimal epistatic networks from integrated sequence and mutational protein data. S. Cocco, L. Posani, R. Monasson. PNAS doi: 10.1073/pnas.2312335121 (2024) Supplementary Material

[2] Learning the differences: a transfer-learning approach to predict antigen immunogenicity and T-cell receptor specificity. B. Bravi, A. Di Gioacchino, J. Fernandez-de-Cossio-Diaz, A. M. Walczak, T. Mora, S. Cocco, and R. Monasson. eLife 12:e85126 (2023)

[3] Machine learning for evolutionary-based and physics-inspired protein design: Current and future synergies. C. Malbranke, D. Bikard, S. Cocco, R. Monasson, J. Tubiana. Current Opinion in Structural Biology 80:102571 (2023)

[4] Mutational paths in protein-sequence landscapes: from sampling to mean-field characterization. E. Mauri, S. Cocco, R. Monasson. Physical Review Letters 130, 158402 (2023)

[5] An evolution-based model for designing chorismate mutase enzymes . W.P. Russ, M. Figliuzzi, C. Stocker, P. Barrat-Charlaix, M. Socolich, P. Kast, D. Hilvert, R. Monasson, S. Cocco, M. Weigt, R. Ranganathan. Science 369, 440-5 (2020)

[6] Learning protein constitutive motifs from sequence data . J. Tubiana, S. Cocco, R. Monasson. eLife 2019;8:e39397 (2019). See also the press release.

Session 1: Algorithms and data structures for sequences

Vizitig: context-rich exploration of sequencing datasets

Bastien DEGARDINS[1], Charles PAPERMAN[1] and Camille MARCHET[1]

1 Univ. Lille, CNRS, Inria, Centrale Lille, UMR 9189 CRIStAL, F-59000 Lille, France

Corresponding author: camille.marchet@univ-lille.fr

Keywords

transcriptomics, pangenomics, de Bruijn graph, visualization, databases

Abstract

Recent advances in k-mer indexing have facilitated the cataloging and rapid querying of planetary-scale genomic data. While these indices excel at high-throughput sequence lookups, they often lack context-rich exploration capabilities and rely on simplistic match-based queries. This gap hinders deeper investigations into variants, regulatory elements, and other features crucial for pangenomic and transcriptomic analyses. We present Vizitig, a novel system that harnesses a de Bruijn graph as the core data structure. By directly encoding overlapping k-mers from both genome and transcriptome data, Vizitig supports the processing of partially or completely unassembled sequences, making it broadly applicable from collections of genomes to eukaryotic RNA-seq. Vizitig integrates k-mer indices into a database framework, providing an intuitive, metadata-aware approach to querying. Users can select candidate regions by specific annotations (e.g., genes, motifs) or sample-specific features (e.g., abundance, presence or absence in annotated genes or samples), retrieving relevant graph neighborhoods and associated metadata from extensive datasets.

REINDEER2: practical abundance index at scale

Yohan HERNANDEZ–COURBEVOIE[1], Mikael SALSON Chloé BESSIERE[1], Haoliang XUE[1], Daniel GAUTHERET[1], Camille MARCHET[1] and Antoine LIMASSET[1]

1 Univ. Lille, CNRS, Centrale Lille, UMR 9189 CRIStAL, F-59000 Lille, France

2 I2BC, Univ. Paris-Saclay, CNRS, CEA, Gif sur Yvette, France

3 IRMB, INSERM U1183, Hopital Saint-Eloi, Universite de Montpellier, Montpellier, France & CRCT, Inserm, CNRS, Univ. Toulouse III-Paul Sabatier, Centre de Recherches en CancÅLerologie de Toulouse, Toulouse, France

4 Independent researcher

Corresponding author: camille.marchet@univ-lille.fr

Keywords

abundance index, k-mer indexing, succinct structures, transcriptomics

Abstract

Over the past decade, significant efforts have been made to develop indexing solutions capable of querying sequence presence in large genomic data repositories. Recent indexing approaches have made giant steps toward the ultimate goal of indexing repositories like the SRA and ENA, leveraging k-mers for efficiency. In the case of indexing RNA samples, querying k-mer abundance is equally important as the presence itself. The currently available methods for indexing abundances either fail to scale to the vast number of datasets, lose variants, or lack precision in abundance estimation. Moreover, the rapid accumulation of sequencing data presents a significant computational challenge for these structures, which are mostly static.

We introduce REINDEER2, a novel k-mer abundance index that addresses these limitations by providing three key properties: scalability, dynamicity, and tunable precision. Unlike recent methods that sacrifice memory for completeness, REINDEER2 indexes all k-mers, ensuring nucleotide-level exploration remains possible. Additionally, it supports high-throughput queries, enabling rapid retrieval of k-mer abundance across large-scale transcriptomic datasets. One of the key advantages of REINDEER2 is its tunable abundance precision. Furthermore, REINDEER2 supports updatability: new datasets can be added efficiently without requiring a complete reindexing process. We report REINDEER2's great efficiency at indexing collections of 1,000–10,000 RNA-seq samples, and demonstrate its capacity to provide abundance estimations comparable to state-of-the-art methods.

Availability: github. com/ Yohan-HernandezCourbevoie/ REINDEER2

OReO: Optimizing Read Order for practical compression

Mathilde GIRARD[1], Lea VANDAMME[1], Bastien CAZAUX[1] and Antoine LIMASSET[1]

1 Univ. Lille, CNRS, Centrale Lille, UMR 9189 CRIStAL, F-59000 Lille, France

Corresponding author: mathilde.girard@univ-lille.fr

Keywords

Compression, Third generation sequencing, Ordering, Assembly

Abstract

Recent advances in high-throughput and third-generation sequencing technologies have created significant challenges in storing and managing the rapidly growing volume of read datasets. Although more than 50 specialized compression tools have been developed, employing methods such as reference-based approaches, customized generic compressors, and read reordering, many users still rely on common generic compressors (e.g., gzip, zstd, xz) for convenience, portability, and reliability, despite their low compression ratios. Here, we introduce OReO, a simple read-reordering framework that achieves high compression performance without requiring specialized software for decompression. By grouping overlapping reads together before applying generic compressors, OReO exploits inherent redundancies in sequencing data and achieves compression ratios on par with state-of-the-art tools. Moreover, because it relies only on standard decompressors, OReO avoids the need for dedicated installations and maintenance, removing a key barrier to practical adoption. We evaluated OReO on both ONT and HiFi genomic datasets of varying sizes and complexities. Our results demonstrate that OReO provides substantial compression gains with comparable resource usage and outperforms dedicated methods in decompression speed. The OReO code is open source and available at github.com/girunivlille/oreo.

Structural Space of Microproteins with Protein Language Models

Simon HERMAN[1], Guillaume BOUVIER[2], Chris PAPADOPOULOS[1], Paul ROGINSKI[1], Olivier LESPINET[1], Anne LOPES[1]

1 Institute for Integrative Biology of the Cell (I2BC), Université Paris-Saclay, CEA, CNRS

2 Structural Bioinformatics Unit, Department of Structural Biology and Chemistry, Institut Pasteur, CNRS UMR3528, C3BI

Corresponding Author: anne.lopes@i2bc.paris-saclay.fr

Keywords

Microproteins, structural properties, folding potential, intergenic ORFs, protein language models

Abstract

Microproteins—defined as proteins comprising fewer than 100 amino acids—have long been overlooked due to challenges in detection arising from their small size and low expression. However, emerging evidence reveals that they are key regulators of translation, development, metabolism, and cellular stress responses, and are implicated in diseases such as cancer and cardiovascular disorders. Recent advances in ribosome profiling and mass spectrometry have uncovered a vast transient microproteome originating from non-canonical open reading frames (ORFs) overlapping annotated genes or located in UTRs or intergenic regions. Although many of these sequences lack signs of selection and are evolutionarily transient, some impact fitness and may represent early stages of de novo gene formation. In this study, we introduce a novel computational framework that employs deep learning–based protein language models (pLMs) to infer the structural properties of microproteins without relying on evolutionary information. Using embeddings from ProtT5-XL, we analyzed thousands of microproteins from annotated sources (e.g., Uniprot) and potential iORF-encoded microproteins across diverse eukaryotic genomes with varying GC content. Using simple dimensionality reduction of these embeddings, we constructed a comprehensive map of the microprotein structural landscape, confirming that amino acid composition and residue ordering are the primary determinants of structure. Then we fine-tuned a classifier that demonstrated robust performance in predicting structural categories, capturing additional signals beyond those revealed by embedding dimensionality reduction. Our results indicate that annotated microproteins occupy narrow, well-defined regions of the structural space, whereas iORF-encoded microproteins exhibit a broader, GC-dependent distribution. Specifically, low-GC iORFs are biased toward encoding transmembrane peptides, while high-GC iORFs predominantly yield disordered proteins. Moreover, certain iORF sequences fall into a

“void” region not populated by canonical proteins, suggesting a region that might have been counterselected by evolution. Taken together, these findings challenge prior knowledge of protein-coding potential and offer fresh insights into the potential evolutionary emergence and structural diversity of microproteins. Our results characterize the structural landscapes of two distinct yet interconnected microproteomes—the annotated coding microproteome and the unannotated, noncoding counterpart—and lay the groundwork for new hypotheses about the molecular evolution of coding sequences from noncoding origins.

GrAnnoT, a tool for efficient and reliable annotation transfer through pangenome graph

Nina Marthe[1], Matthias Zytnicki[2], and Francois Sabot[1]

DOI not yet assigned

Abstract

The increasing availability of genome sequences has highlighted the limitations of using a single reference genome to represent the diversity within a species. Pangenomes, encompassing the genomic information from multiple genomes, offer thus a more comprehensive representation of intraspecific diversity. However, pangenomes in form of graph often lack annotation information, which limits their utility for forward analyses. We introduce here GrAnnoT, a tool designed for efficient and reliable annotation transfer using such graphs, by projecting existing annotations from a source genome to the graph and subsequently to other embedded genomes. GrAnnoT was benchmarked against state-of-the-art tools on pangenome graphs and linear genomes from rice, human, and *E. coli*. The results demonstrate that GrAnnoT is consensual, conservative, and fast, outperforming alignment-based methods in accuracy or speed or both. It provides informative outputs, such as presence-absence matrices for genes, and alignments of transferred features between source and target genomes, aiding in the study of genomic variations and evolution. GrAnnoT's robustness and replicability across different species make it a valuable tool for enhancing pangenome analyses. GrAnnoT is available under the GNU GPLv3 licence at `https://forge.ird.fr/diade/dynadiv/grannot`.

Keywords: Pangenome, graph, annotation transfer

[1]DIADE unit, UM, Cirad, IRD, 911 avenue Agropolis Montpellier F-34391, France, [2]MIAT, INRAE, Auzeville-Tolosane F31320, France

Correspondence
nina.marthe@ird.fr,francois.sabot@ird.fr

Introduction

Recent advances in genome sequencing and assembly have allowed access to a massive and increasing number of genome sequences per species. This has highlighted the fact that a single individual is not enough to represent the whole diversity of a species. Indeed, while currently prevalent, the use of a single reference genome has been shown to bias some analysis (Chen et al., 2021; Martiniano et al., 2020; Maurstad et al., 2024). This has led to the development of the concept of pangenomics across the whole tree of life (Bayer et al., 2020; Liao et al., 2023; Miga and Wang, 2021; Rouli et al., 2015; Shi et al., 2023; Tranchant-Dubreuil et al., 2019). A pangenome aims to represent the complete genomic information from several genomes of the same species or group, in order to better represent the intra-specific/group diversity. This information can be organized in different ways depending on the type of study involved. The pangenome graph structure has recently emerged as a solution to store and model these pangenomes. This structure has the advantage of containing the whole sequence information (genic and inter-genic regions) and of encoding the relationships between the genomes (which regions are identical, where are the variations) in a compact and comprehensive way. These pangenome graphs can be stored in variation graphs in GFA format, the standardized text format used by many tools. It is human-readable and represents the multiple genome alignment. Multiple forms of graphs that can be stored in the GFA format; the graphs we consider in this paper are bi-directed and acyclic genome graphs (see 2.1 for details on the graphs used).

Pangenome graphs have already proven their usefulness to better understand the structure and dynamics of genomes, for structural variations detection, or for genotyping for instance (Rice et al., 2023; Zhou et al., 2022). However, their practical use still has limitations, as the tools to manipulate them are often still in development. In particular, these graphs usually do not contain any annotation information. Similarly to a genome without annotation, a pangenome graph without embedded biological information is less useful, and the variations present in the graph are harder to interpret. Therefore any variation found in the graph has to be reported back to a single linear annotated genome to see if it overlaps with a region of interest. An annotated graph would then allow to more easily study the structure and evolution of a species pangenome.

Furthermore, despite the recent appearance of AI-based tools (Holst et al., 2023), *de novo* genome annotation is a long and complicated process. However, good quality and manually curated annotations already exist for many species, usually for a single linear individual genome. These existing annotations can be transferred to other non-annotated genomes to add meaningful biological information. This transfer operation is much faster than a *de novo* annotation, requiring far less computation and resources. Annotation transfer between linear genomes is usually performed using a blast-like approach (e.g. Liftoff (Shumate and Salzberg, 2021)), by mapping/aligning the sequence of the annotated elements on the target genome. However, this approach is not currently adapted to pangenome graphs. Indeed, alignment on a graph is more complex than alignment on a linear sequence, and existing graph alignment tools are quite recent and still under maturation. However, the graph itself already represents a global alignment between the annotated genome and the other embedded genomes. Thus, this alignment can be used to project the coordinates of the annotated elements from the linear genome to the graph, and any information concerning a linear genome position can be transferred to the graph. Once the annotation is transferred onto the graph, it can be transferred back to any other genome

embedded in the graph. Therefore, any information added to the graph will benefit all the embedded genomes.

We developed GrAnnoT (Graph Annotation Transfer) to perform these operations. This command-line tool will allow to gather more information in pangenome graphs and better harness the many graphs that have been already produced (Rice et al., 2023; Shi et al., 2023; Zhou et al., 2022). It includes functions to transfer annotations and to study the variations present in the graph in the annotated elements. We applied it to different pangenome graphs to ensure it works with various species, and compared it to existing methods for annotation transfer. Compared to these methods, GrAnnoT is consensual, conservative, and fast. It also offers informative outputs to allow the user to review the transfers performed, such as a presence-absence matrix, alignments between the genes in different genomes or a list of the variations found in annotated regions.

1. Implementation

GrAnnoT is implemented in Python 3.10, as a Linux command-line tool that can be installed as a standard python package. It uses the package tqdm and the external program bedtools (Quinlan, 2014) (that must be accessible in the user or in the global path). The code is available on the IRD forge (`https://forge.ird.fr/diade/dynadiv/grannot`) under the GNU GPLv3 licence.

1.1. Code overview

GrAnnoT performs annotation transfer from an annotated genome (the source genome) to a pangenome graph (Figure 1). It can also transfer the annotation from the graph to one, several or all other genomes embedded in the graph (the target genomes). It takes as input a pangenome graph in GFA 1.1 format (which includes the source and target genomes), and the annotation of the source genome in GFF3 format. For the sake of clarity, in the present paper the graph considered has a unique path for each haplotype, but the method works for assemblies with several chromosomes/contigs.

The annotation transfer only relies on the graph structure, harnessing the multiple alignment it naturally represents. GrAnnoT projects the coordinates between the graph and the genomes, transferring annotations in an alignment-free manner.

Once the annotation has been loaded, GrAnnoT outputs the graph annotation in GAF format, which describes the paths of the annotated features in the graph. It can then output the annotation in GFF format of a chosen set of target linear genomes included in the graph. These transfers can be filtered through sequence identity and coverage scores. For these transfers, the alignment of each feature between the source genome and the target one can be outputted in a Clustal-like format, as well as a list of all the variations recorded in the alignments. Finally, a presence-absence matrix for gene features summarizes the transfer on the target genomes.

1.2. Implementation details

The first step is to find the start and stop positions of each node from the graph on the source and target genomes (Figure 1, step 1). For that, GrAnnoT follows the paths of these genomes in the graph and computes the start and stop positions of the nodes for each of them; these positions are then stored in BED files. Then, the BED file representing the source genome is compared to its annotation file using bedtools intersect (Quinlan, 2014). The resulting BED file is processed to compute the paths of the features in the graph and output the graph annotation in GAF format.

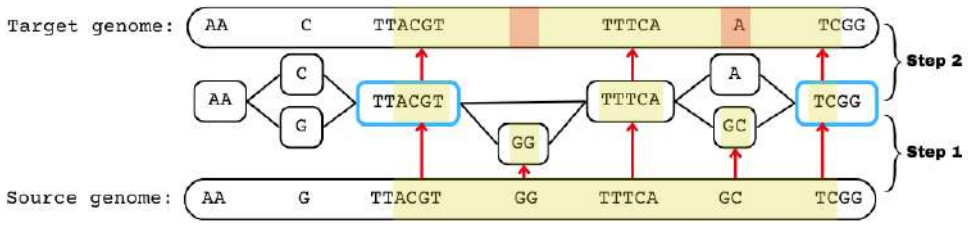


Figure 1 – GrAnnoT overview. Step 1: the position of the feature is projected from the source genome to the graph using the positions of the nodes on the source genome. Step 2: the position of the feature is projected from the graph to the target genome. The first and last nodes from the feature that are on the target genome (the blue ones) are the ends of the feature in this genome, and everything in between is considered as part of the feature. The differences between the two genomes in this region in terms of path in the graph mirror the differences between the two versions of the feature.

In order to transfer an annotation to a target genome, the sub-path of the genome corresponding to the feature is extracted (Figure 1, step 2). For that, all the nodes from the original feature path are looked for in the target genome path. These nodes are then grouped into copies of the feature, and for each copy the first and the last nodes are considered as the ends of the feature's copy in the target genome. All the nodes between them in the target genome path are expected to be part of the feature's copy to transfer, including the nodes absent from the original feature path, corresponding to insertions. Nodes from the original feature path that are not found in the target genome correspond to deletions. An insertion and a deletion at the same locus in the graph correspond to a substitution.

For the transfer itself, only the two nodes at the ends of the feature path on the target genome are considered (nodes in blue in Figure 1). The BED file previously computed reporting the positions of the nodes on the target genome is used to locate these two nodes on the genome.

Transferred features are then filtered based on the coverage (in base) and the identity level between the source and the target genomes. These parameters are estimated by computing the cumulated length of the shared and different nodes between the paths of the features in the two genomes. The output is finally printed out in the GFF format.

If the user is interested in the differences between the source and target annotation, GrAnnoT can provide a detailed comparison between the feature alternative paths in the source and any embedded target genome. For that, it can output the variations details in a text format that describes all the variations present in the feature (node deletion, insertion, substitution). A Clustal-like alignment file of all the transferred features based on their alternative paths is similarly generated.

2. Benchmark

2.1. Data and tools for benchmarks

The main test data used in this paper is a rice pangenome graph built with 13 genomes (Kawahara et al., 2013a; Zhou et al., 2020) using minigraph-cactus v2.8.2 with default options (Hickey et al., 2023; see supplementary data for the exact commands) and the cv Nipponbare as reference. The rice genome is 380-410Mb long and has 12 chromosomes. The annotation used

as source (Kawahara et al., 2013b) includes 57,585 *gene* features for 813,790 total features, and is rich in transposable elements (15,848/57,858≈27%).

GrAnnoT was also tested on a graph of the human chromosome 1 with 92 haplotypes (from Liao et al., 2023) and an *E. coli* 12 genomes (Jangir et al., 2022) graph built using the same protocol as for rice (detailed commands available online, Marthe and Sabot, 2025b).

GrAnnoT was compared to existing and state-of-the-art tools (see below) that can also perform annotation transfer in order to assess its efficiency, and using the different data presented before to test its replicability and robustness. All analyses were ran on a biprocessor Intel(R) Xeon(R) CPU E5-2650 v4 @ 2.20GHz with 48 HT CPU computer with 144Gb of RAM, under RockyLinux 9.1 Blue Onyx.

The current state-of-the-art annotation transfer tool for linear genome sequences is Liftoff (Shumate and Salzberg, 2021). It is widely used (Alonge et al., 2022; Kim et al., 2021; Wang et al., 2021; Yang et al., 2023), and relies primarily on the alignment of the nucleic sequences of the annotated features from the source genome upon the target one. However, since Liftoff does not use a pangenome graph to transfer annotations, the comparison with GrAnnoT is biased by the graph itself, whose structure partially impacts the results of GrAnnoT transfer (see below).

Liftoff approach to transfer annotations can be adapted to a pangenome graph by aligning the sequences of the annotated features to the graph. Graph pangenome alignment tools can be thus compared to GrAnnoT for graph annotation transfer: GraphAligner was chosen for this purpose (Rautiainen and Marschall, 2020), as a state-of-the-art tool for aligning long sequences on a graph.

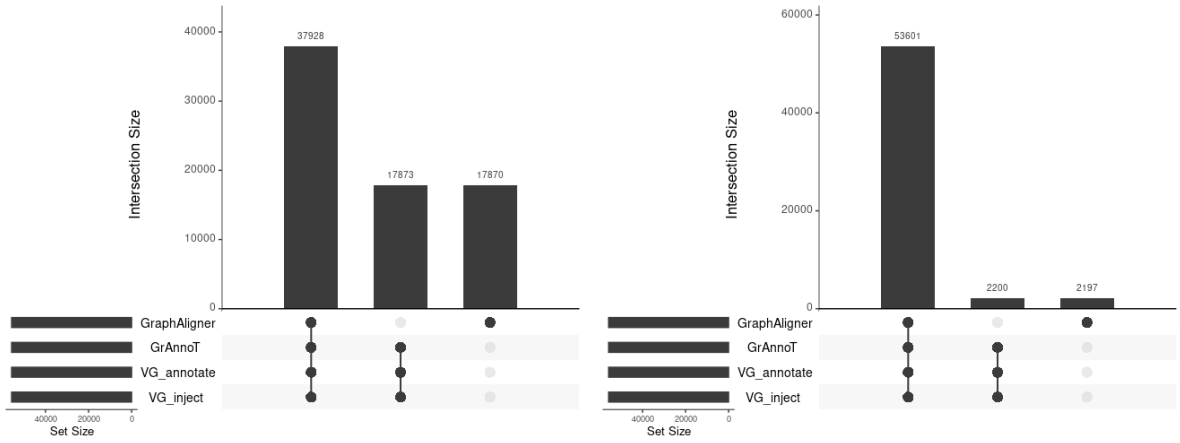
VG and ODGI are state-of-the-art tools for pangenome graph manipulation (Garrison et al., 2018; Guarracino et al., 2022). They do not have options specifically designed to transfer annotations between genomes of the graph, but they do have options to project coordinates between the graph and its embedded genomes, specifically *odgi position* and *vg inject/surject*. These options can be adapted to transfer annotations between genomes, and from a genome to the graph. VG also has an option for annotation transfer on the graph (*vg annotate*). The results and execution time of all these functions were compared to GrAnnoT.

The versions of the tools used are available in the supplementary data. The complete exact commands used for those benchmark are available online (Marthe and Sabot, 2025b) The Jupyter notebooks used for the analysis are available on our Forge (`https://forge.ird.fr/diade/dynadiv/grannot`, Marthe et al., 2025). All the data used for the analysis and the outputs are available online (Marthe and Sabot, 2025a,b).

2.2. Comparison of the transfers

Comparison with other tools

Results were evaluated for the two types of transfers that GrAnnoT can perform: from genome to graph and from genome to genome. In both cases, the transfer was performed with the different tools described before when possible. Then, for each transferred feature, its positions provided by the different tools were compared. Given a feature, we consider two transfers as different if they placed the feature at different positions. A transfer is specific to a tool if it is different from all the other transfers. By definition, a feature transfer is also specific to a tool if the feature is only transferred by this tool.



(a) Features transferred at the exact same position **(b)** Features transferred at a distance of 1 bp maximum

Figure 2 – Genome to graph transfer comparison, Upset representation. Each vertical bar represents the number of identical transfers between the different tools specified below the bar. Two transfers are considered identical if they placed the feature at the exact same path in the graph and either at the exact same position on the nodes (a) or at a distance of maximum 1 nucleotide (b).

We tested GrAnnoT, GraphAligner and VG (*inject* and *annotate* functions) by transferring the annotation of the cv Nipponbare (Kawahara et al., 2013b) to the rice pangenome graph. We tested GrAnnoT, Liftoff, VG (*inject+surject* functions) and ODGI (*position* function) by transferring the annotation of cv Nipponbare to cv Azucena.

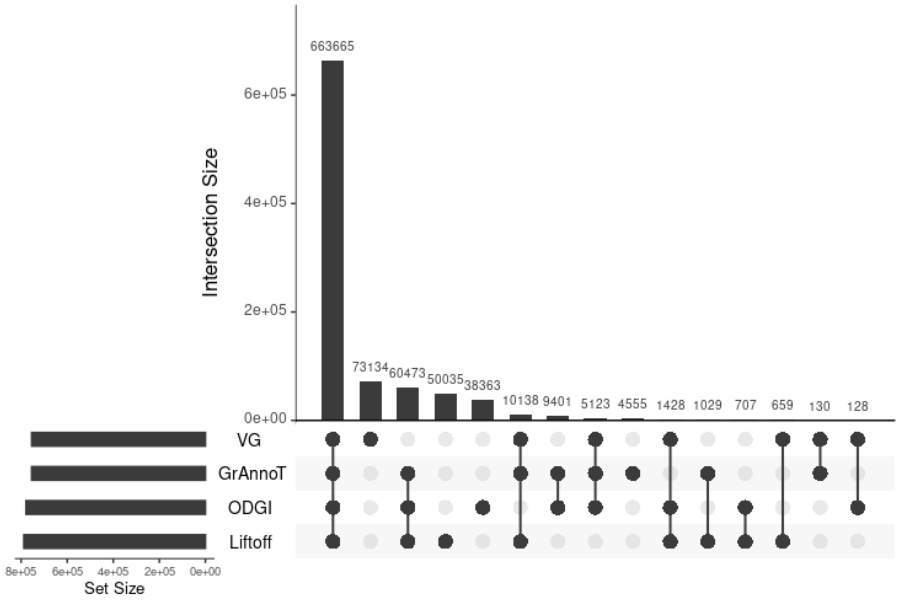
Genome to graph transfer:

The three methods that do not perform alignment (GrAnnoT, *vg inject* and *vg annotate*) have the exact same results for all the features. For ~32% of the features transferred by GraphAligner (17,870 features out of 55,798), the output is different from the other tools (Figure 2a). However, when allowing a difference of 1 bp on the position on the path, ~88% of the GraphAligner-specific transfers (15,673 out of 17,870 transfers) are then considered identical to the transfers from the other tools (Figure 2b). Further verification showed that these 1 bp differences from GraphAligner are alignment errors, where 1 bp is missing in 5' or 3' in the transferred feature sequence. Such differences are minor and acceptable for certain applications, but not in the context of annotation. Because of that, the current version of GraphAligner does not seem to be suitable for annotation transfer.

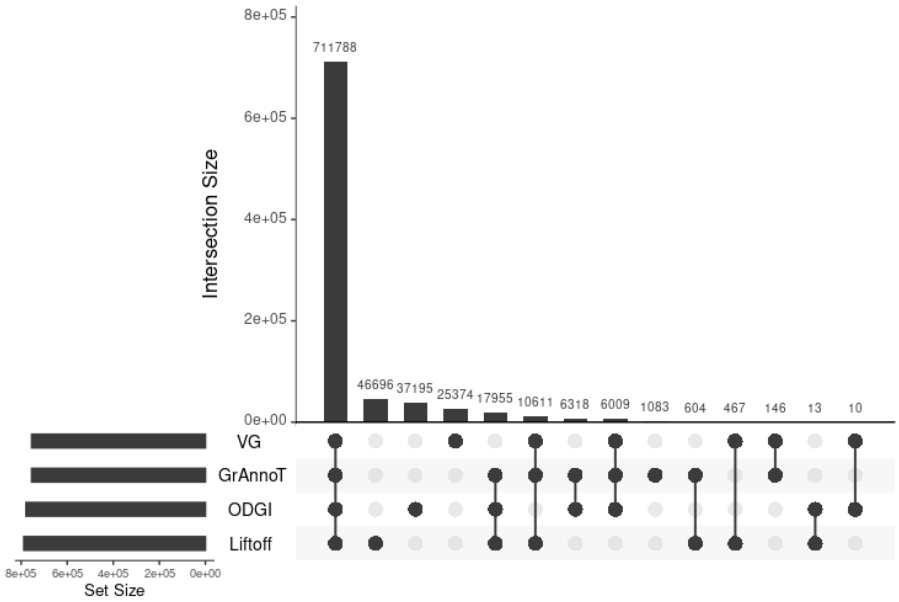
Genome to genome transfer:

Most of the transfers between genomes are identical between the four tools (663,665/918,973≈72%). GrAnnoT seems to be the most consensual tool as it has the least specific transfers (Figure 3a) compared to the other tools.

When looking at the tool-specific transfers, VG stands out the most, with 73,134 specific transfers. However, when allowing a difference of 10 bp between the transfers, VG has ~65.3% less specific transfers. Some of these VG specific transfers were manually compared to the transfers from the other tools for the same feature, and were identified as errors from VG (see supplementary Figure S10 for an example). The 10bp difference tolerance revealed Liftoff and ODGI as the most divergent tools (with 46,696 and 37,195 specific transfers, respectively; Figure 3b).



(a) Features transferred at the exact same position



(b) Features transferred at a distance of 10 bp maximum

Figure 3 – Genome to genome transfer comparison, Upset representation. Each vertical bar represents the number of identical transfers between the different tools specified below the bar. Two transfers are considered identical if they placed the feature either at the exact same positions on the target genome (a) or at a distance of maximum 10 nucleotide (b).

Regarding the Liftoff-specific transfers, most are features that only Liftoff can transfer. Indeed, ~56% of them are inter-chromosomal translocations, *i.e.* features that are on a different chromosome between the source and the target genome (Figure 4). These transfers cannot be performed with GrAnnoT, VG or ODGI, as graphs are currently built chromosome-per-chromosome to reduce complexity, and therefore cannot represent such events. Thus, features on different chromosomes between Nipponbare and Azucena cannot be transferred by any of the graph-based approaches, and are found only by Liftoff.

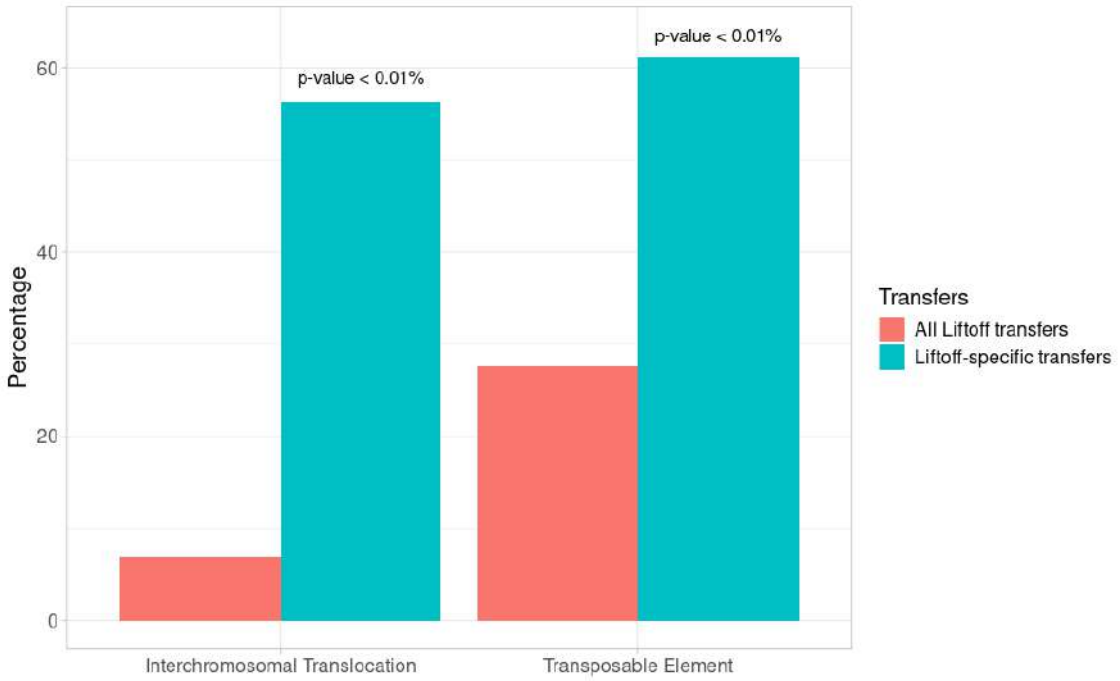


Figure 4 – Transposable element and inter-chromosomal translocation percentages in all Liftoff transfer *vs* Liftoff-specific transfers. The Liftoff-specific transfers are enriched in translocations and in transposable elements compared to all the other Liftoff transfers. Detailed data and *p*-value calculation are available in supplementary data (table S6 and S7).

Furthermore, when the annotations of these Liftoff specific-transferred features were thoroughly looked at, it appeared that they are enriched in transposable elements (TE) (*p*-value < 0.01%; Figure 4). This could explain why so many elements are on different chromosomes between the two varieties, since transposable elements are mobile in the genome and can jump between chromosomes (Hayward and Gilbert, 2022; Wicker et al., 2007). The Liftoff-specific transfers that are on the same chromosome are also enriched in transposable elements (*p*-value < 0.01%; Figure 5), as their ability to move in the genome makes them often not syntenic: encoding the relationships between such elements in the graph with the current pangenome graph tools still seems complex. Since these relationships are not correctly encoded by the graph, the TE annotation transfer cannot be reliably performed by tools such as GrAnnoT, which only uses the structure of the graph.

Most of the ODGI-specific transfers place a feature on a very small interval on the target genome. For instance, among the 37,195 ODGI-specific transfers, ~65% of the features (24,058) are placed on an interval of length 0 nucleotide, and ~30% (11,320) on an interval of length 1 nucleotide. These transfers should be discarded, as they are of no biological meaning in terms of genes.

Robustness

Back and forth transfer: Two consecutive transfers with Liftoff and GrAnnoT allowed to compare how conservative these tools are. The first transfer was performed from Nipponbare to Azucena with the two tools. Then, the resulting Azucena GFF was used as source to perform the second

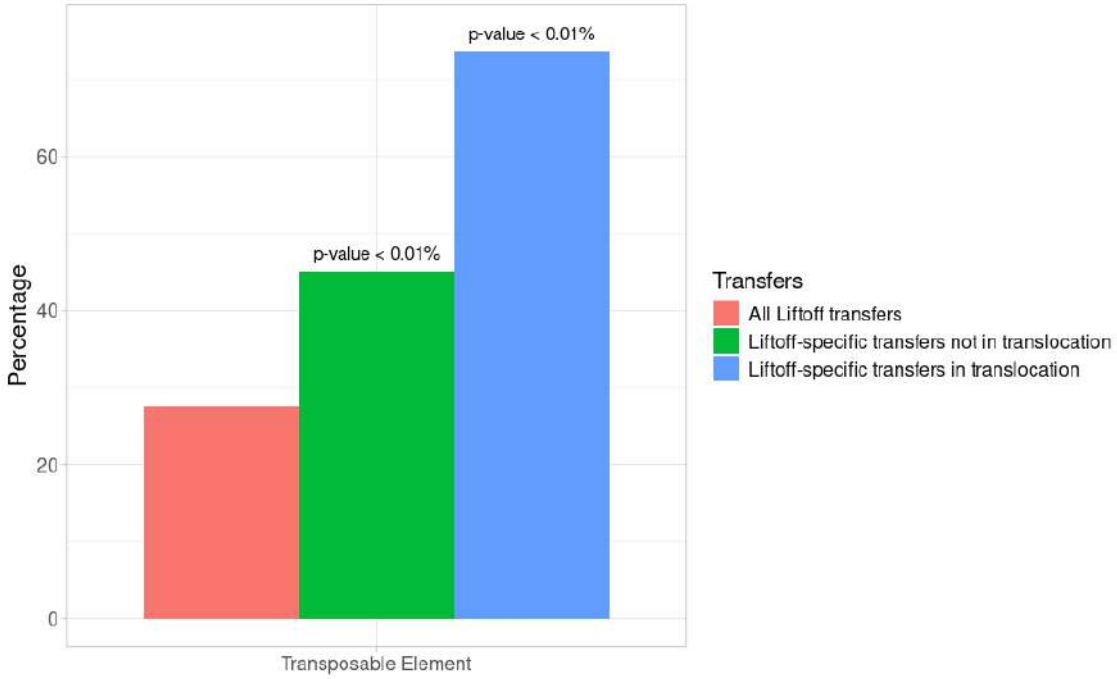


Figure 5 – Transposable element percentages in all Liftoff transfer *vs* Liftoff-specific transfers. Compared to the other Liftoff transfers, Liftoff-specific transfers are enriched in transposable elements, whether or not they are in a translocation. Detailed data and *p*-value calculation are available in supplementary data (table S8 and S9).

transfer, from Azucena back to Nipponbare. The resulting GFF for Nipponbare was compared to its first original annotation to measure the loss of information during these transfers.

Liftoff loses less features during the two-round process (table 1). This can be explained by the fact that Liftoff is better at finding non-syntenic features and handles interchromosomal translocations, as shown previously. However, while GrAnnoT did not loose any more annotation on the way back to the original sequence, Liftoff lost an additional 257 of them. In addition, when comparing the positions of the features before and after the two transfers, GrAnnoT shows better results that Liftoff, with only 3.8% of the features being located at a different position compared to the original annotation, *versus* 10.4% of discrepancies for Liftoff. In addition, after manual verification, it appeared that the features misplaced by GrAnnoT in the second transfer are features where an extremity was shortened during the first transfer due to a deletion. Thus, the feature transferred during the second transfer was incomplete regarding the true annotation, but the transfer itself occurred correctly.

Finally, some features found by Liftoff are placed on a different chromosome than the original, as the transfer is alignment-based only and does not rely on synteny. In this regard GrAnnoT is more conservative than Liftoff. Indeed, orthologous copies are sometimes considered to guarantee a better conservation of gene function compared to paralogous copies, according to the ortholog conjecture (Nevers et al., 2020; Rogozin et al., 2014). As the graph conserves the synteny, GrAnnoT is more likely to transfer annotations between orthologous copies than between paralogous copies.

Impact of the reference genome for graph construction: The graphs used were built with Minigraph-Cactus, which requires a reference genome as anchor, that can thus bias the graph structure (Andreace et al., 2023). To test the replicability of the GrAnnoT approach, transfers through two

	GrAnnoT	Liftoff
Loss in first transfer	7,961	1,622
Loss in second transfer	0	257
Total loss	7,961	1,879
Same position	47,184	48,482
Different position	841	5,625
1-10bp difference	393	360
11-100bp difference	275	648
101-1000bp difference	165	776
>1000bp difference	8	1,210
Different chromosome	0	2,631
Total transfers	48,025	54,107

Table 1 – GrAnnoT and Liftoff comparison on back and forth transfer. The input annotation for first transfer included 55,986 features. The loss corresponds to the number of features not transferred in either transfer (Nipponbare to Azucena or Azucena to Nipponbare). The other rows show how many features were at the same or at different positions before and after the two transfers.

	Nipponbare reference	Natel Boro reference
Total transfers	48,025	45,946
Loss	7,961	10,040
Specific transfers	2,376	297
	Comparison between the two graphs	
Common transfers	45,256	
Different transfers	393	
1-10bp difference	169	
11-100bp difference	87	
101-1000bp difference	76	
>1000bp difference	61	
Different chromosome	0	

Table 2 – Comparison of GrAnnoT transfers using graphs with different reference genomes. The input annotation for the transfer included 55,986 features. The loss corresponds to the number of features not transferred. The other rows show how many features were placed at the same or at different positions when transferred with the two graphs.

different graphs were compared. The two graphs have the same genomes embedded, but a different reference genome to initiate the graph. The reference genomes used for the two graphs are the annotated genome IRGSP-1.0 (Nipponbare), and Os127652RS1 (Natel Boro) (Zhou et al., 2020). Annotation transfer from genome to genome was performed with these two graphs, and the positions of the common transferred features were compared.

Among the 48,322 features transferred on Azucena, 2,673 (~5.5%) were not transferred by both graphs. Among the 45,649 features transferred by both graphs, only 393 (~0.9%) were not transferred at the same location (table 2).

The amount of features not transferred by both graphs is not negligible, but it can be explained by the choice of the reference genome for the graph construction, Natel Boro. Indeed, among the 11 genomes in the graph that are not involved in the transfer (not Nipponbare or Azucena), Natel Boro is among the furthest genetically speaking, as shown in the phylogenetic tree in the genomes original paper (Zhou et al., 2020). Thus, it makes sense that the graph centered

	Rice		Human		E.coli	
Total features to transfer	55,986		282,668		9,467	
Features transferred by GrAnnoT	48,025	85.78%	276,681	97.88%	7,230	76.37%
GrAnnoT-specific transfers	72	0.13%	4,647	1.64%	81	0.86%
Features transferred by Liftoff	54,363	97.10%	275,249	97.38%	7,940	83.87%
Liftoff-specific transfers	6,410	11.45%	3,264	1.12%	790	8.34%
Features transferred by both tools	47,951	85.65%	258,092	91.31%	7,146	75.48%
Same position	46,431	82.93%	256,927	90.89%	6,789	71.71%
Different position	1,520	2.71%	1,165	0.41%	357	3.77%
1-10bp difference	795	1.42%	623	0.22%	216	2.28%
11-100bp difference	284	0.51%	150	0.05%	60	0.63%
101-1000bp difference	155	0.28%	46	0.02%	18	0.19%
>1000bp difference	164	0.29%	346	0.12%	63	0.67%
Different chromosome	122	0.22%	0	0%	0	0%
Runtime Liftoff	00:23:45		00:10:27		00:00:08	
Runtime GrAnnoT	00:08:11		00:14:47		00:00:20	

Table 3 – Comparison between GrAnnoT and Liftoff in several species. Each feature in the input annotation was transferred using GrAnnoT and Liftoff. When the feature has been transferred by both tools, the two positions given were compared to see how different they are.

around Nipponbare displays better performance for annotation transfer from Nipponbare. This showcases the importance of the choice of the reference genome for the graph construction, that must be adapted to the use case of the graph. In the case of annotation transfer, using the annotated source genome is a good solution.

Comparison with other species: GrAnnoT was compared to Liftoff using two other datasets: a pangenome graph of the human chromosome 1 (Liao et al., 2023) and an *E. coli* pangenome graph (Jangir et al., 2022). Both of these graphs were made with Minigraph-Cactus. For the rice graph, the transfer was again made from Nipponbare to Azucena; for the human graph, the transfer was made from CHM13 to GrCH38; for the *E.coli* graph, the transfer was made from K_12_MG1655_09949b0 to O127_H6_E2348_69_193637c. These comparisons checked if the positions of the features transferred by both approach are consistent, to see if the results observed in the rice pangenome graph were replicable with graphs from other type of dataset/organisms/phylum.

It appears that for the three species, most of the features are transferred by both tools (~85.7% for rice, ~91.3% for human and ~75.5% for *E. coli*) (table 3). Additionally, a large part of these features are placed at the exact same position by Liftoff and GrAnnoT (~96.8% for rice, ~99.6% for human and ~95% for *E. coli*). As expected, some features are transferred only by Liftoff, but for the human graph GrAnnoT-specific transfers appear in negligible quantities. This better transfer capacity for the two tool in human may be due to the lesser diversity of human genomes compared to rice (mean 15.6 millions SNP for 64 human haplotypes *vs* 9.4 millions for only 16 rice ones, respectively; Ebert et al., 2021; Wei et al., 2024), and and even more so compared to *E. coli*. In addition, the annotation of human genes is probably better curated than in rice, with less hypothetical genes that may be false positive, also explaining the better transfer for both tools on human reference.

	GrAnnoT	Liftoff	VG	ODGI
Run time	00:08:11	00:23:45	07:26:54	70:23:41

Table 4 – Run time comparison for genome to genome transfer in rice.

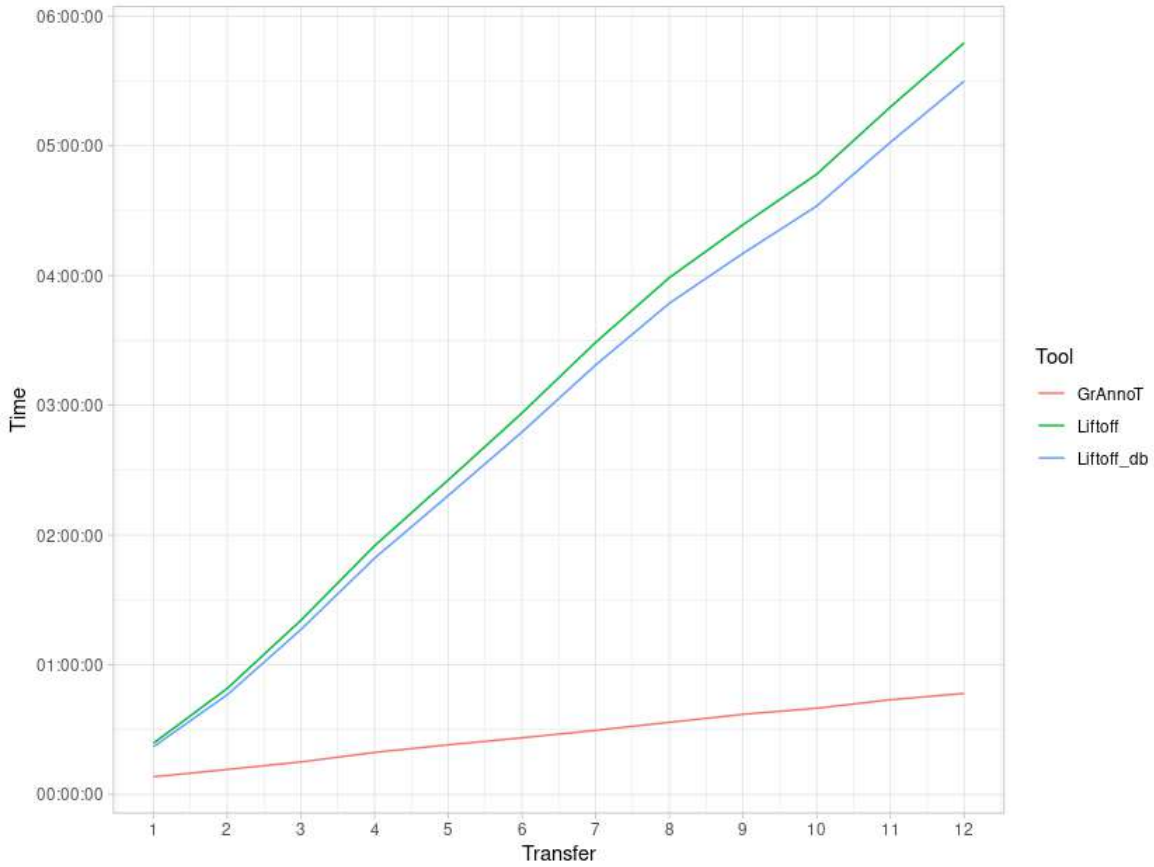


Figure 6 – Genome to genome transfer comparison. GrAnnoT and Liftoff run time for 1-12 transfers were measured using the command */usr/bin/time*. Liftoff was run both in GFF and DB mode. Detailed time points are available in supplementary data in table S10.

2.3. Time and memory usage results

The execution time for the transfer from genome to genome with the different tools was measured using the command */usr/bin/time* (table 4). The results show that GrAnnoT has the best run time, and that ODGI and VG are substantially slower that GrAnnoT and Liftoff.

GrAnnoT was further compared to Liftoff in terms of run time and memory usage.

Several transfers were performed with both tools to compare the run times, because GrAnnoT is advantaged when multiple transfers are requested, as GrAnnoT starts by pre-processing the graph and loading the graph annotation. These steps only need to be done once, no matter how many annotation transfers to target genomes are performed.

Liftoff can be run in GFF mode or in database mode; the database mode needs less time since the GFF annotation file has already been processed. Both of these mode were compared to GrAnnoT.

The results show that GrAnnoT is faster than Liftoff to perform one annotation transfer (~8 minutes *vs* ~22 minutes), and even more to perform twelve (~47 minutes *vs* ~5 hours and 30

minutes, see Figure 6). However, this comparison doesn't take into account the time needed to build the graph. When adding the graph construction time (~4h55mn on our infrastructure) to the GrAnnoT 12 transfers time, we still get a duration (~5h41mn) equivalent to Liftoff transfers (~5h47min or ~5h29min). Additionally, GrAnnoT can give supplementary informative output that describe the transfers performed, such as a presence-absence matrix or alignment files of the transferred features.

For the human graph, GrAnnoT is not faster than Liftoff for one transfer (see the last lines of table 3). However, as shown on Figure 6, for several transfers GrAnnoT is more advantageous. We tested the runtime of GrAnnoT for the annotation transfer on 10 haplotypes, and got ~45 minutes in total. This is significantly lower than the time for one transfer multiplied by 10 (~1h44min), which is what we can expect of 10 Liftoff transfers from the results in Figure 6.

3. Applications

To assess the use of GrAnnoT annotation transfer, in particular the informative outputs complementary to the GFF itself, we analyzed a few characteristics of the annotation transfers between the Nipponbare and Azucena cultivars. More precisely, we verified that the variations in the graph reported by GrAnnoT are distributed as biologically expected, in a way that does not disrupt the proteins coded by the gene features.

3.1. Indel rate in different feature types

We looked at the positions of the indel variations (insertion or deletion) in the different feature types that correspond to different parts of the genes. These variations are expected to be less present in the CDS compared to the rest of the gene due to selection pressure, because the resulting changes in the coded protein are more important.

The feature types that were compared are:

- the whole gene feature itself
- the mRNA
- the 5'UTR
- the exons
- the CDS
- the introns
- the 3'UTR

These feature types have different average lengths, with the gene being the longest element (since it contains all the others), and the CDS being shorter than the exon summed size, for instance. This induces a bias in the number of indel found by feature type; if the indels are randomly distributed, we expect more indels in the feature type that has the longest cumulated length. To counter this bias, for each feature type we reported the number of indels found to its cumulated length, obtaining the average number of indel per position.

The results displayed in Figure 7 show that, as expected, the CDS have the fewest indels and the non-coding regions (UTR and introns) have the most. This confirms that the variations in the graph reported by GrAnnoT are consistent with the current understanding of genome variation selection.

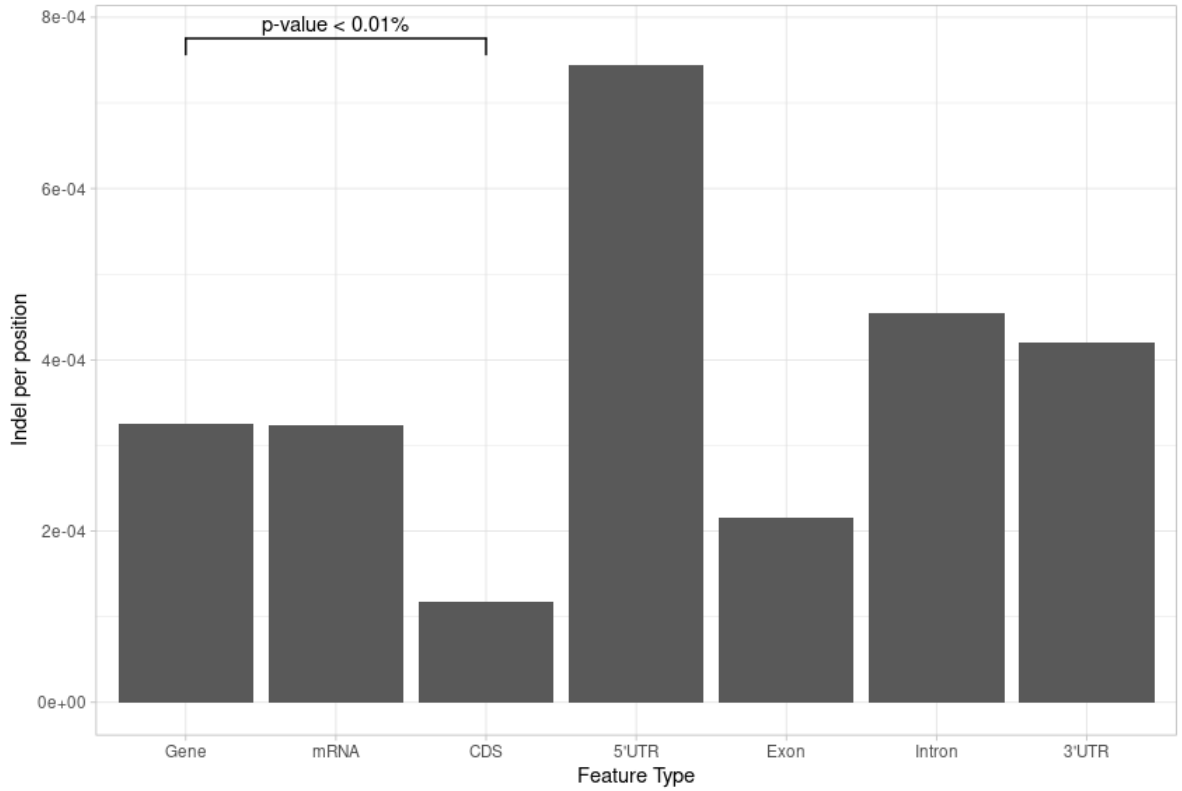


Figure 7 – Indel distribution. Each bar represents the number of indels (insertion or deletion) per position in the corresponding feature type. As expected, the CDS are the most conserved and thus have the least indels, and the non-coding regions (UTR and introns) are the least conserved and have the most indels. Detailed data and *p*-value calculation are available in supplementary data (table S11).

3.2. Frameshit mutations in different feature types

Indels can modify the protein coded by a gene, but indels in CDS are particularly impactful when they change the reading frame. We calculated the rate of frameshift mutations (indel whose length is not a multiple of 3) among the indels, for all feature types. We expect to have a lower ratio of frameshift mutations in the CDS compared to the non-coding regions, because of the selection pressure.

The results displayed in Figure 8 show that the CDS have the lowest percentage of frameshift variation from their indels, and that the introns have the highest.

3.3. Substitutions position in different feature types

The substitutions are smaller variations than the indels, so they are expected to have a smaller impact. However their distribution in CDS is not expected to be uniform. Indeed, substitutions on the third position of a codon is more likely to be silent than a substitution on the two other positions. Because of that, in CDS the third codon position usually has more substitutions than the two other positions (Sanchez et al., 2005).

On Figure 9, we show that the CDS indeed has more substitutions on the third codon position than the other two positions, while the other gene elements have more homogeneous substitution distributions.

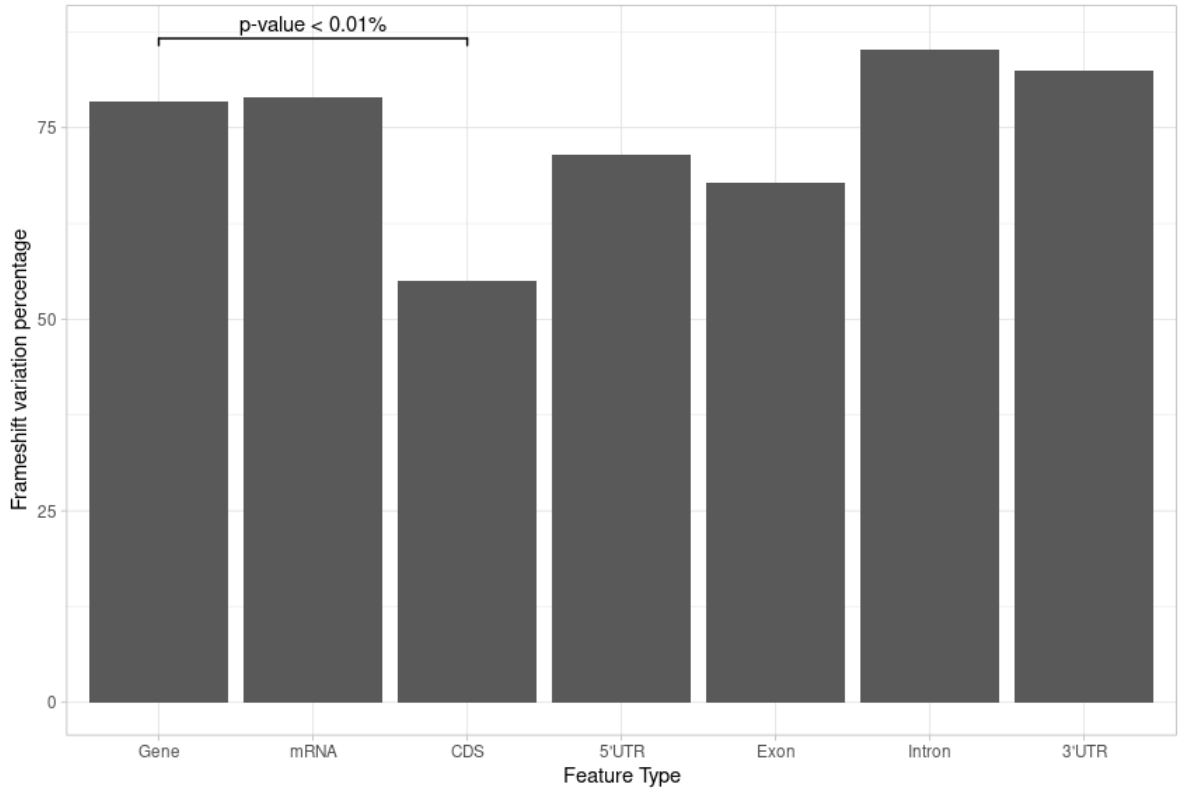


Figure 8 – Frameshift indel distribution. Each bar represent the percentage of frameshift variation (length not multiple of 3) among all the indels in each feature type. As expected, the CDS have the least frameshift variation, since these variations impact significantly the protein coded. Detailed data and *p*-value calculation are available in supplementary data (table S12).

To find the codon positions we had to take into account the splicing of the mRNA. The CDS elements in the annotation only correspond to a fraction of the real CDS in the mRNA. Thus the substitution positions are not relative to the real CDS, and finding the third position of the codon required to add the context of the preceding CDS fragments. Thus adjustment was only done for the CDS elements in the annotation, since they were the only element of interest. This explains why the exons do not follow the CDS tendency in Figure 9, contrary to Figures 7 and 8.

3.4. Pangene set analysis

The PAV matrix output was computed with all the genomes in the rice graph (minus the source genome Nipponbare), and was used to compute the core, dispensable and shell gene set from the Nipponbare cv in this pangenome.

We found ~58% of core genes and ~33% of dispensable genes (table 5) in our pangenome graph, which is similar to what is seen in the literature when accounting for the different threshold chosen in each study (with ~53-62% of core and ~38% of dispensable gene families for instance; Wang et al., 2018).

Conclusion

In the present study, we presented the first tool able to efficiently transfer annotation on a pangenome graph from one of its embedded genomes and reverse, GrAnnoT. It relies on the

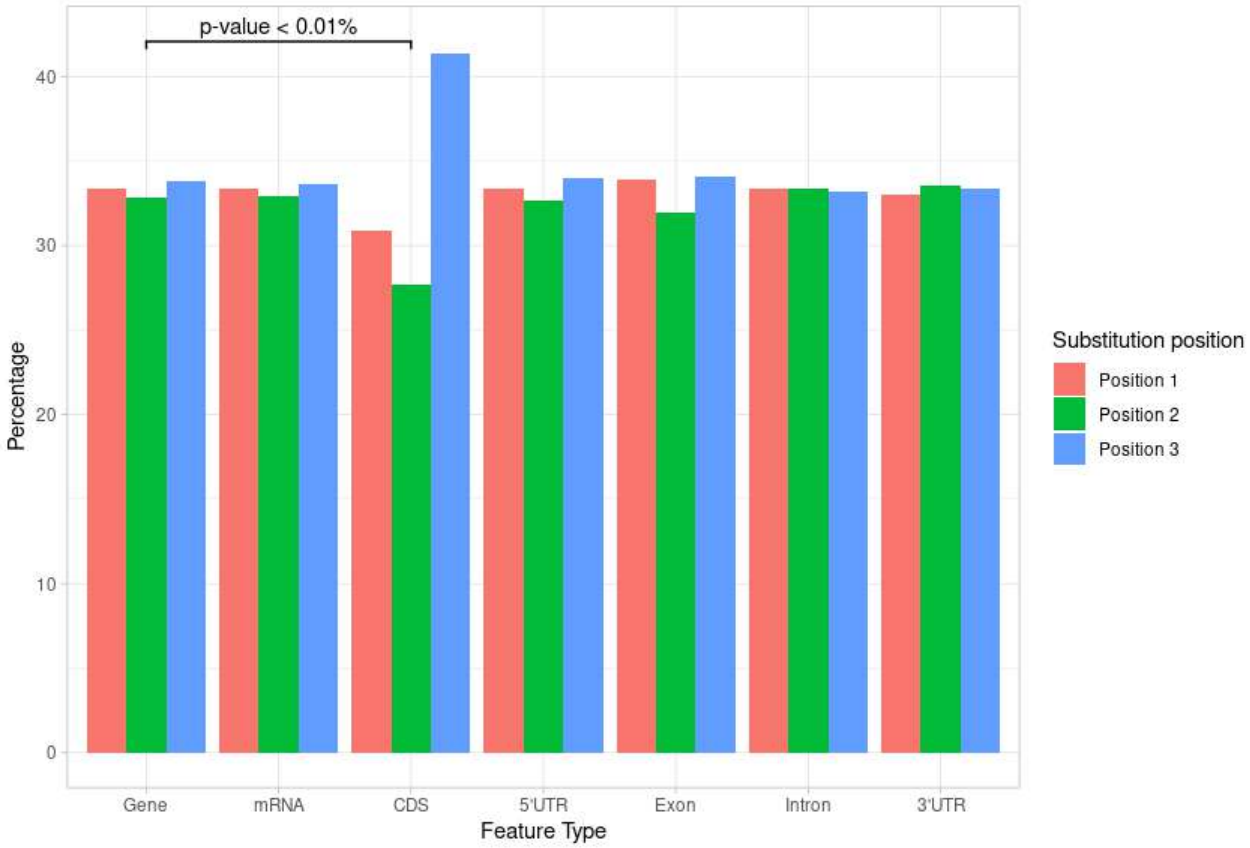


Figure 9 – Substitution positions. For each feature type, the percentage of substitutions that are on each of the three codon positions is displayed. In the CDS, the third position has more substitutions than the two other positions. For the other feature types, we don't see that the positions multiple of 3 have more substitutions than the others. Detailed data and *p*-value calculation are available in supplementary data (table S13).

	Core genes	Dispensable genes	Shell genes
Presence percentage	100% - 95%	95% - 10%	10% - 0%
Number of genes	32,537	18,403	5 046
Percentages of genes	58.1%	32.9%	9%

Table 5 – Core, dispensable and shell gene set. The population size is 12, and there are 55,986 genes in total.

already performed alignment that created the graph. We benchmarked GrAnnoT on rice and human pangenomes, and showed that it is fast, reliable and efficient, compared to state-of-the-art tools for linear genomes. It is a robust, replicable tool working on any type of species for which a pangenome graph is available. In addition, GrAnnoT can provide useful outputs, such as the alignments of the gene sequence between source and target, or a presence/absence matrix. In the near future, we plan to optimize the transfer time through parallelization, and to implement the inference of the impact of mRNA and CDS mutations on the resulting proteins.

Acknowledgements

The authors want to thank Sebastien Ravel from PHIM/CIRAD for his valuable discussions and help on packaging.

The authors acknowledge the ISO 9001 certified IRD i-Trop HPC (South Green Platform) at IRD Montpellier for providing HPC resources that have contributed to the research results reported within this paper. URL: https://bioinfo.ird.fr/- http://www.southgreen.fr

Fundings

Nina Marthe is supported by a PhD funding from the Agence Nationale de la Recherche as part of the France2030 program under the reference "ANR-22-PEAE-4".

Conflict of interest disclosure

The authors declare that they comply with the PCI rule of having no financial conflicts of interest in relation to the content of the article. The authors declare the following non-financial conflict of interest: FS is a PCI recommender. The LLM agent LeChat was used to prepare the abstract section once the manuscript finished.

Data, script, code, and supplementary information availability

Data and results are available online (data: https://doi.org/10.23708/DO1RTF,Marthe and Sabot, 2025a, results: https://doi.org/10.23708/RRSKRA, Marthe and Sabot, 2025b). Script and codes are available online (https://doi.org/10.23708/TW3KYV, Marthe et al., 2025).

Supplementary data

Data and tools used :

- Data
  - *E.coli* graph (built with data from Jangir et al., 2022)
  - Human graph (from Liao et al., 2023)
  - Rice graph (built with data from Kawahara et al., 2013a; Zhou et al., 2020)
- Tools
  - minigraph-cactus v2.8.2 (Hickey et al., 2023)
  - Liftoff v1.6.3 (Shumate and Salzberg, 2021)
  - ODGI v0.8.6-11-ga1f169cc (Guarracino et al., 2022)
  - VG v1.58.0 (Garrison et al., 2018)
  - GraphAligner Branch master commit daec67f67a2f50d648a6aa30cbbe5a2949583061 (Rautiainen and Marschall, 2020)

	Different chromosome	Same chromosome	*P*-value
Liftoff-specific transfers	3604	2806	< 0.01%
Other Liftoff transfers	122	47831	

Table S6 – Interchromosomal translocation rates in Liftoff transfers. The *p*-value measures the enrichment in interchromosomal translocations in the Liftoff-specific transfers, and was computed with Pearson's Chi-squared test.

```
LOC_Os01g01050                  ACAAGTCACAGGGAGGAGTC    20
GrAnnoT_Lifotff_ODGI_transfer   ACAAGTCACAGGGAGGAGTC    20
VG_transfer                     ACAAGTCACAGGGAGGAGTC    20
                                ********************

                            ...
                            ...
                            ...
LOC_Os01g01050                  TCTAT--------CTATCTA    512
GrAnnoT_Lifotff_ODGI_transfer   tctatctatctatctatcta    520
VG_transfer                     tctatctatctatctatcta    520
                                *****        *******

                            ...
                            ...
                            ...
LOC_Os01g01050                  TATACATGACGATATGATCC    4131
GrAnnoT_Lifotff_ODGI_transfer   TATACATGACGATATGATCC    4139
VG_transfer                     TATACATGACGA--------    4131
                                ************
```

Figure S10 – Extract of the alignment of gene LOC_Os01g01050 and its transfers in Azucena by different tools. VG transfer appears to have an error as the positions it gives miss the last 8 bases of the gene. The gene total length is conserved in VG transfer because there is an insertion in Azucena in the middle of the gene.

	Transposable elements	Other features	*P*-value
Liftoff-specific transfers	3919	2491	< 0.01%
Other Liftoff transfers	11053	36900	

Table S7 – Transposable elements rates in Liftoff transfers. The *p*-value measures the enrichment in transposable elements in the Liftoff-specific transfers, and was computed with Pearson's Chi-squared test.

	Transposable elements	Other features	*P*-value
Liftoff-specific transfers on the same chromosome	1263	1543	< 0.01%
All Liftoff transfers	13709	46410	

Table S8 – Transposable elements rates in Liftoff transfers. The *p*-value measures the enrichment in transposable elements in the Liftoff-specific transfers on the same chromosome, and was computed with Pearson's Chi-squared test.

	Transposable elements	Other features	*P*-value
Liftoff-specific transfers on a different chromosome	2656	948	< 0.01%
All Liftoff transfers	12316	38443	

Table S9 – Transposable elements rates in Liftoff transfers. The *p*-value measures the enrichment in transposable elements in the Liftoff-specific transfers in interchromosomal translocations, and was computed with Pearson's Chi-squared test.

	GrAnnoT	Liftoff GFF	Liftoff DB
1 transfer	00:08:11.64	00:23:45	00:22:08
2 transfers	00:11:36.00	00:49:03	00:46:07
3 transfers	00:15:04.47	01:20:34	01:16:18
4 transfers	00:19:29.69	01:55:02	01:49:17
5 transfers	00:22:59.03	02:25:29	02:18:16
6 transfers	00:26:13.64	02:56:24	02:47:37
7 transfers	00:29:41.71	03:29:06	03:18:38
8 transfers	00:33:23.48	03:59:03	03:47:11
9 transfers	00:37:06.03	04:23:31	04:10:16
10 transfers	00:39:54.23	04:46:52	04:32:09
11 transfers	00:43:50.98	05:17:58	05:01:38
12 transfers	00:46:45.57	05:47:34	05:29:53

Table S10 – GrAnnoT and Liftoff time comparison for 1-12 transfers

	Positions without indel	Positions with indel	*P*-value
Gene	141980198	46148	< 0.01%
CDS	74201787	8762	

Table S11 – Indel rates in genes and CDS. Annotations were transferred between cv Nipponbare and Azucena, and the number of insertions and deletions was analyzed. The *p*-value measures the enrichment in indel in gene features, and was computed with Pearson's Chi-squared test.

	Non-frameshift indel	Frameshift indel	*P*-value
Gene	9959	36189	< 0.01%
CDS	3950	4812	

Table S12 – Frameshift indel rates in genes and CDS. Annotations were transferred between cv Nipponbare and Azucena, and the insertions and deletions lengths were analyzed. The *p*-value measures the enrichment in indel causing a frameshift in gene features, and was computed with Pearson's Chi-squared test.

	Substitutions on position 1 or 2	Substitutions on position 3	*P*-value
Gene	171763	87638	< 0.01%
CDS	73005	51562	

Table S13 – Substitution positions in nucleotide triplets in genes and CDS. Annotations were transferred between cv Nipponbare and Azucena, and the substitution positions were analyzed. The *p*-value measures the enrichment in substitutions on position 3 of the nucleotide triplets in CDS features, and was computed with Pearson's Chi-squared test.

References

Alonge M, Lebeigle L, Kirsche M, Jenike K, Ou S, Aganezov S, Wang X, Lippman ZB, Schatz MC, Soyk S (2022). *Automated assembly scaffolding using RagTag elevates a new tomato system for high-throughput genome editing*. *Genome Biology* **23**. https://doi.org/10.1186/s13059-022-02823-7. URL: http://dx.doi.org/10.1186/s13059-022-02823-7.

Andreace F, Lechat P, Dufresne Y, Chikhi R (2023). *Comparing methods for constructing and representing human pangenome graphs*. *Genome Biology* **24**. https://doi.org/10.1186/s13059-023-03098-2. URL: http://dx.doi.org/10.1186/s13059-023-03098-2.

Bayer PE, Golicz AA, Scheben A, Batley J, Edwards D (2020). *Plant pan-genomes are the new reference*. *Nature Plants* **6**, 914–920. https://doi.org/10.1038/s41477-020-0733-0. URL: http://dx.doi.org/10.1038/s41477-020-0733-0.

Chen NC, Solomon B, Mun T, Iyer S, Langmead B (2021). *Reference flow: reducing reference bias using multiple population genomes*. *Genome Biology* **22**. https://doi.org/10.1186/s13059-020-02229-3. URL: http://dx.doi.org/10.1186/s13059-020-02229-3.

Ebert P, Audano PA, Zhu Q, Rodriguez-Martin B, Porubsky D, Bonder MJ, Sulovari A, Ebler J, Zhou W, Mari RS, Yilmaz F, Zhao X, Hsieh P, Lee J, Kumar S, Lin J, Rausch T, Chen Y, Ren J, Santamarina M, et al. (2021). *Haplotype-resolved diverse human genomes and integrated analysis of structural variation*. *Science* **372**, eabf7117. https://doi.org/10.1126/science.abf7117. eprint: https://www.science.org/doi/pdf/10.1126/science.abf7117. URL: https://www.science.org/doi/abs/10.1126/science.abf7117.

Garrison E, Sirén J, Novak AM, Hickey G, Eizenga JM, Dawson ET, Jones W, Garg S, Markello C, Lin MF, Paten B, Durbin R (2018). *Variation graph toolkit improves read mapping by representing genetic variation in the reference*. *Nature Biotechnology* **36**, 875–879. https://doi.org/10.1038/nbt.4227. URL: http://dx.doi.org/10.1038/nbt.4227.

Guarracino A, Heumos S, Nahnsen S, Prins P, Garrison E (2022). *ODGI: understanding pangenome graphs*. *Bioinformatics* **38**. Ed. by Peter Robinson, 3319–3326. https://doi.org/10.1093/bioinformatics/btac308. URL: http://dx.doi.org/10.1093/bioinformatics/btac308.

Hayward A, Gilbert C (2022). *Transposable elements*. *Current Biology* **32**, R904–R909.

Hickey G, Monlong J, Ebler J, Novak AM, Eizenga JM, Gao Y, Abel HJ, Antonacci-Fulton LL, Asri M, Baid G, Baker CA, Belyaeva A, Billis K, Bourque G, Buonaiuto S, Carroll A, Chaisson MJP, Chang PC, Chang XH, Cheng H, et al. (2023). *Pangenome graph construction from genome alignments with Minigraph-Cactus*. *Nature Biotechnology* **42**, 663–673. https://doi.org/10.1038/s41587-023-01793-w. URL: http://dx.doi.org/10.1038/s41587-023-01793-w.

Holst F, Bolger A, Günther C, Maß J, Triesch S, Kindel F, Kiel N, Saadat N, Ebenhöh O, Usadel B, Schwacke R, Bolger M, Weber AP, Denton AK (2023). *Helixer–de novoPrediction of Primary Eukaryotic Gene Models Combining Deep Learning and a Hidden Markov Model*. https://doi.org/10.1101/2023.02.06.527280. URL: http://dx.doi.org/10.1101/2023.02.06.527280.

Jangir PK, Yang Q, Shaw LP, Caballero JD, Ogunlana L, Wheatley R, Walsh T, MacLean RC (2022). *Pre-existing chromosomal polymorphisms in pathogenic E. coli potentiate the evolution of resistance to a last-resort antibiotic*. *eLife* **11**. https://doi.org/10.7554/elife.78834. URL: http://dx.doi.org/10.7554/eLife.78834.

Kawahara Y, Bastide M, Hamilton JP, Kanamori H, McCombie WR, Ouyang S, Schwartz DC, Tanaka T, Wu J, Zhou S, Childs KL, Davidson RM, Lin H, Quesada-Ocampo L, Vaillancourt B,

Sakai H, Lee SS, Kim J, Numa H, Itoh T, et al. (2013a). *Improvement of the Oryza sativa Nipponbare reference genome using next generation sequence and optical map data*. *Rice* **6**. https://doi.org/10.1186/1939-8433-6-4. URL: https://doi.org/10.1186/1939-8433-6-4.

Kawahara Y, Bastide M, Hamilton JP, Kanamori H, McCombie WR, Ouyang S, Schwartz DC, Tanaka T, Wu J, Zhou S, Childs KL, Davidson RM, Lin H, Quesada-Ocampo L, Vaillancourt B, Sakai H, Lee SS, Kim J, Numa H, Itoh T, et al. (2013b). *Improvement of the Oryza sativa Nipponbare reference genome using next generation sequence and optical map data*. *Rice* **6**. https://doi.org/10.1186/1939-8433-6-4. URL: http://dx.doi.org/10.1186/1939-8433-6-4.

Kim BY, Wang JR, Miller DE, Barmina O, Delaney E, Thompson A, Comeault AA, Peede D, D'Agostino ER, Pelaez J, Aguilar JM, Haji D, Matsunaga T, Armstrong EE, Zych M, Ogawa Y, Stamenković-Radak M, Jelić M, Veselinović MS, Tanasković M, et al. (2021). *Highly contiguous assemblies of 101 drosophilid genomes*. *eLife* **10**. https://doi.org/10.7554/elife.66405. URL: http://dx.doi.org/10.7554/eLife.66405.

Liao WW, Asri M, Ebler J, Doerr D, Haukness M, Hickey G, Lu S, Lucas JK, Monlong J, Abel HJ, Buonaiuto S, Chang XH, Cheng H, Chu J, Colonna V, Eizenga JM, Feng X, Fischer C, Fulton RS, Garg S, et al. (2023). *A draft human pangenome reference*. *Nature* **617**, 312–324. https://doi.org/10.1038/s41586-023-05896-x. URL: http://dx.doi.org/10.1038/s41586-023-05896-x.

Marthe N, Sabot F (2025a). *Data for GrAnnoT*. Version V1. https://doi.org/10.23708/DO1RTF. URL: https://doi.org/10.23708/DO1RTF.

Marthe N, Sabot F (2025b). *Output for Grannot*. Version V1. https://doi.org/10.23708/RRSKRA. URL: https://doi.org/10.23708/RRSKRA.

Marthe N, Sabot F, Zytnicki M (2025). *GrAnnoT source code and scripts*. Version V1. https://doi.org/10.23708/TW3KYV. URL: https://doi.org/10.23708/TW3KYV.

Martiniano R, Garrison E, Jones ER, Manica A, Durbin R (2020). *Removing reference bias and improving indel calling in ancient DNA data analysis by mapping to a sequence variation graph*. *Genome Biology* **21**. https://doi.org/10.1186/s13059-020-02160-7. URL: http://dx.doi.org/10.1186/s13059-020-02160-7.

Maurstad MF, Hoff SNK, Cerca J, Ravinet M, Bradbury I, Jakobsen KS, Præbel K, Jentoft S (2024). *Reference genome bias in light of species-specific chromosomal reorganization and translocations*. https://doi.org/10.1101/2024.06.28.599671. URL: http://dx.doi.org/10.1101/2024.06.28.599671.

Miga KH, Wang T (2021). *The Need for a Human Pangenome Reference Sequence*. *Annual Review of Genomics and Human Genetics* **22**, 81–102. https://doi.org/10.1146/annurev-genom-120120-081921. URL: http://dx.doi.org/10.1146/annurev-genom-120120-081921.

Nevers Y, Defosset A, Lecompte O (2020). *Orthology: Promises and Challenges*. In: *Evolutionary Biology—A Transdisciplinary Approach*. Springer International Publishing, pp. 203–228. https://doi.org/10.1007/978-3-030-57246-4_9. URL: http://dx.doi.org/10.1007/978-3-030-57246-4_9.

Quinlan AR (2014). *BEDTools: The Swiss-Army Tool for Genome Feature Analysis*. *Current Protocols in Bioinformatics* **47**. https://doi.org/10.1002/0471250953.bi1112s47. URL: http://dx.doi.org/10.1002/0471250953.bi1112s47.

Rautiainen M, Marschall T (2020). *GraphAligner: rapid and versatile sequence-to-graph alignment*. *Genome Biology* **21**. https://doi.org/10.1186/s13059-020-02157-2. URL: http://dx.doi.org/10.1186/s13059-020-02157-2.

Rice ES, Alberdi A, Alfieri J, Athrey G, Balacco JR, Bardou P, Blackmon H, Charles M, Cheng HH, Fedrigo O, Fiddaman SR, Formenti G, Frantz LAF, Gilbert MTP, Hearn CJ, Jarvis ED, Klopp C, Marcos S, Mason AS, Velez-Irizarry D, et al. (2023). *A pangenome graph reference of 30 chicken genomes allows genotyping of large and complex structural variants*. *BMC Biology* **21**. https://doi.org/10.1186/s12915-023-01758-0. URL: http://dx.doi.org/10.1186/s12915-023-01758-0.

Rogozin IB, Managadze D, Shabalina SA, Koonin EV (2014). *Gene Family Level Comparative Analysis of Gene Expression in Mammals Validates the Ortholog Conjecture*. *Genome Biology and Evolution* **6**, 754–762. https://doi.org/10.1093/gbe/evu051. URL: http://dx.doi.org/10.1093/gbe/evu051.

Rouli L, Merhej V, Fournier PE, Raoult D (2015). *The bacterial pangenome as a new tool for analysing pathogenic bacteria*. *New Microbes and New Infections* **7**, 72–85. https://doi.org/10.1016/j.nmni.2015.06.005. URL: http://dx.doi.org/10.1016/j.nmni.2015.06.005.

Sanchez R, Morgado E, Grau R (2005). *Gene algebra from a genetic code algebraic structure*. *Journal of Mathematical Biology* **51**, 431–457. https://doi.org/10.1007/s00285-005-0332-8. URL: http://dx.doi.org/10.1007/s00285-005-0332-8.

Shi J, Tian Z, Lai J, Huang X (2023). *Plant pan-genomics and its applications*. *Molecular Plant* **16**, 168–186. https://doi.org/10.1016/j.molp.2022.12.009. URL: http://dx.doi.org/10.1016/j.molp.2022.12.009.

Shumate A, Salzberg SL (2021). *Liftoff: accurate mapping of gene annotations*. *Bioinformatics* **37**. Ed. by Alfonso Valencia, 1639–1643. https://doi.org/10.1093/bioinformatics/btaa1016. URL: http://dx.doi.org/10.1093/bioinformatics/btaa1016.

Tranchant-Dubreuil C, Rouard M, Sabot F (2019). *Plant Pangenome: Impacts on Phenotypes and Evolution*. https://doi.org/10.1002/9781119312994.apr0664. URL: http://dx.doi.org/10.1002/9781119312994.apr0664.

Wang B, Yang X, Jia Y, Xu Y, Jia P, Dang N, Wang S, Xu T, Zhao X, Gao S, Dong Q, Ye K (2021). *High-Quality Arabidopsis Thaliana Genome Assembly with Nanopore and HiFi Long Reads*. *Genomics, Proteomics & Bioinformatics* **20**, 4–13. https://doi.org/10.1016/j.gpb.2021.08.003. URL: http://dx.doi.org/10.1016/j.gpb.2021.08.003.

Wang W, Mauleon R, Hu Z, Chebotarov D, Tai S, Wu Z, Li M, Zheng T, Fuentes RR, Zhang F, Mansueto L, Copetti D, Sanciangco M, Palis KC, Xu J, Sun C, Fu B, Zhang H, Gao Y, Zhao X, et al. (2018). *Genomic variation in 3, 010 diverse accessions of Asian cultivated rice*. *Nature* **557**, 43–49. https://doi.org/10.1038/s41586-018-0063-9. URL: http://dx.doi.org/10.1038/s41586-018-0063-9.

Wei X, Chen M, Zhang Q, Gong J, Liu J, Yong K, Wang Q, Fan J, Chen S, Hua H, Luo Z, Zhao X, Wang X, Li W, Cong J, Yu X, Wang Z, Huang R, Chen J, Zhou X, et al. (2024). *Genomic investigation of 18,421 lines reveals the genetic architecture of rice*. *Science* **385**, eadm8762. https://doi.org/10.1126/science.adm8762. eprint: https://www.science.org/doi/pdf/10.1126/science.adm8762. URL: https://www.science.org/doi/abs/10.1126/science.adm8762.

Wicker T, Sabot F, Hua-Van A, Bennetzen JL, Capy P, Chalhoub B, Flavell A, Leroy P, Morgante M, Panaud O, et al. (2007). *A unified classification system for eukaryotic transposable elements*. *Nature reviews genetics* **8**, 973–982.

Yang C, Zhou Y, Song Y, Wu D, Zeng Y, Nie L, Liu P, Zhang S, Chen G, Xu J, Zhou H, Zhou L, Qian X, Liu C, Tan S, Zhou C, Dai W, Xu M, Qi Y, Wang X, et al. (2023). *The complete and fully-phased diploid genome of a male Han Chinese*. *Cell Research* **33**, 745–761. https://doi.org/10.1038/s41422-023-00849-5. URL: http://dx.doi.org/10.1038/s41422-023-00849-5.

Zhou Y, Zhang Z, Bao Z, Li H, Lyu Y, Zan Y, Wu Y, Cheng L, Fang Y, Wu K, Zhang J, Lyu H, Lin T, Gao Q, Saha S, Mueller L, Fei Z, Städler T, Xu S, Zhang Z, et al. (2022). *Graph pangenome captures missing heritability and empowers tomato breeding*. *Nature* **606**, 527–534. https://doi.org/10.1038/s41586-022-04808-9. URL: http://dx.doi.org/10.1038/s41586-022-04808-9.

Zhou Y, Chebotarov D, Kudrna D, Llaca V, Lee S, Rajasekar S, Mohammed N, Al-Bader N, Sobel-Sorenson C, Parakkal P, Arbelaez LJ, Franco N, Alexandrov N, Hamilton NRS, Leung H, Mauleon R, Lorieux M, Zuccolo A, McNally K, Zhang J, et al. (2020). *A platinum standard pan-genome resource that represents the population structure of Asian rice*. *Scientific Data* **7**. https://doi.org/10.1038/s41597-020-0438-2. URL: http://dx.doi.org/10.1038/s41597-020-0438-2.

Facilitating genome annotation using ANNEXA and long-read RNA sequencing

Nicolaï HOFFMANN[1], Aurore BESSON[1], Edouard CADIEU[1,2], Matthias LORTHIOIS[1,2], Victor LE BARS[1,2], Armel HOUEL[1], Christophe HITTE[1], Catherine ANDRÉ[1], Benoit HÉDAN[1] and Thomas DERRIEN[1,2]

[1] Canine Genetics Team, CNRS, Univ Rennes – UMR6290, IGDR (Institut de Génétique et Développement de Rennes), 35043 Rennes, France

[2] IGDRion plateform, CNRS, Univ Rennes – UMR6290, IGDR (Institut de Génétique et Développement de Rennes), 35043 Rennes, France

Corresponding author: nicolai.hoffmann@univ-rennes.fr, thomas.derrien@univ-rennes.fr

Keywords Genome Annotation, long-read sequencing, lncRNAs, comparative transcriptomics

Abstract *With the advent of complete genome assemblies, genome annotation has become essential for the functional interpretation of genomic data. Long-read RNA sequencing (LR-RNAseq) technologies have significantly improved transcriptome annotation by enabling full-length transcript reconstruction for both coding and non-coding RNAs. However, challenges such as transcript fragmentation and incomplete isoform representation persist, highlighting the need for robust quality control (QC) strategies. This study presents an updated version of ANNEXA, a pipeline designed to enhance genome annotation using LR-RNAseq data while also providing QC for reconstructed genes and transcripts. ANNEXA integrates two transcriptome reconstruction tools, StringTie2 and Bambu, applying stringent filtering criteria to improve annotation accuracy. It also incorporates deep learning models to evaluate transcription start sites (TSSs) and employs the tool FEELnc for the systematic annotation of long non-coding RNAs (lncRNAs). Additionally, the pipeline offers intuitive visualizations for comparative analyses of coding and non-coding repertoires. Benchmarking against multiple reference annotations revealed distinct patterns of sensitivity and precision for both known and novel genes and transcripts and mRNAs and lncRNAs. To demonstrate its utility, ANNEXA was applied in a comparative oncology study involving LR-RNAseq of two human and eight canine cancer cell lines. The pipeline successfully identified novel genes and transcripts across species, expanding the catalog of protein-coding and lncRNA annotations in both species. Implemented in Nextflow for scalability and reproducibility, ANNEXA is available as an open-source tool: https://github.com/IGDRion/ANNEXA.*

Introduction

With the increasing availability of entire genome assemblies, i.e. telomere-to-telomere (T2T), one challenge in genome research is to move from improving genome completeness to refining genome annotation. High-quality genome sequences now enable a more precise characterisation of genes and transcripts, especially in repetitive regions of the genomes, making transcriptome-based annotation a critical step in the functional interpretability of the genomes [1]. To this end, RNA sequencing (RNA-seq) plays a central role in this process by providing direct transcript-level evidence, essential for defining gene structures, alternative splicing events, and non-coding RNA repertoires.

While short-read RNAseq (SR-RNASeq) has shown some limitations to reconstruct full-length transcripts [2], long-read RNA sequencing (LR-RNAseq), provided by platforms such as Pacific Biosciences

(PacBio) and Oxford Nanopore Technologies (ONT), has significantly advanced transcriptome annotation by providing reads that span repeats and also by allowing direct connectivity between distant exons of the same isoform [3]. However, despite these advantages, LR-RNAseq-based transcriptome reconstruction still remains prone to artifacts such as transcript fragmentation and incomplete isoform representation [4]. To ensure accurate genome annotation, robust quality control (QC) strategies are needed to evaluate and refine transcriptome reconstructions.

Several tools have been developed to assemble long reads from LR-RNAseq data. A recent benchmark study from the LRGASP consortium has compared fourteen transcriptome reconstruction and quantification tools [5] and showed that choosing the best program depends on the biological context of the study and the completeness of the reference annotation. Among the benchmarked tools, Bambu [6] and StringTie [7] consistently demonstrated strong performance across multiple metrics. However, one limitation of this benchmark is that it did not explicitly assess tool performance with respect to RNA biotypes, particularly distinguishing between protein-coding transcripts (mRNAs) and long non-coding RNAs (lncRNAs). Given the complexity and heterogeneity of transcriptomes, this distinction could be important since different RNA biotypes may vary in expression levels, structural features, and evolutionary conservation, all of which can impact reconstruction accuracy. This consideration is especially important in light of the substantial expansion of the human lncRNA catalogue in recent Gencode releases [8], underscoring the need for tools that can accurately reconstruct and annotate both known and novel transcripts across diverse RNA classes.

To address these limitations, we present ANNEXA, a novel pipeline designed to extend reference annotation based on LR-RNASeq data and assess the quality of novel model transcripts. ANNEXA integrates two transcriptome reconstruction tools, Bambu [6] and StringTie [7], and provides users with stringent filters to improve annotation accuracy. It also incorporates a deep learning strategy to potentially remove incomplete transcript models by evaluating transcription start sites (TSSs) of all novel transcripts. ANNEXA is also designed to systematically annotate and evaluate the annotation of lncRNAs by incorporating the FEELnc program [9]. It provides intuitive visual representations of the annotation, facilitating comparative analysis of coding and non-coding repertoires. To illustrate the usability of ANNEXA in a comparative oncology project, we sequenced two human and eight canine cancer cell lines from mucosal melanomas, histiocytic sarcomas and osteosarcoma using ONT direct cDNA sequencing, and identified novel human and canine genes/transcripts, some of which being conserved in the two species.

By implementing a structured framework for quality control and annotation, ANNEXA enhances the reliability of long-read transcriptomics, ensuring more comprehensive and biologically meaningful extended genome annotations.

Methods

Overview of ANNEXA

ANNEXA is a pipeline that extends user-provided reference annotations with novel genes and transcript isoforms from long-read sequencing data. It only uses three parameter files: a reference genome, a reference annotation and mapping files (**Fig.1.A**). Unlike technology-specific pipelines such as nf-core/isoseq [10], which is tailored for PacBio data, ANNEXA is compatible with both Nanopore

and PacBio RNA sequencing technologies. The pipeline is organized into four main modules, described below:

1. Transcriptome reconstruction
2. Coding potential evaluation and transcript classification
3. Transcript filtering and full-length assessment
4. Quality Control

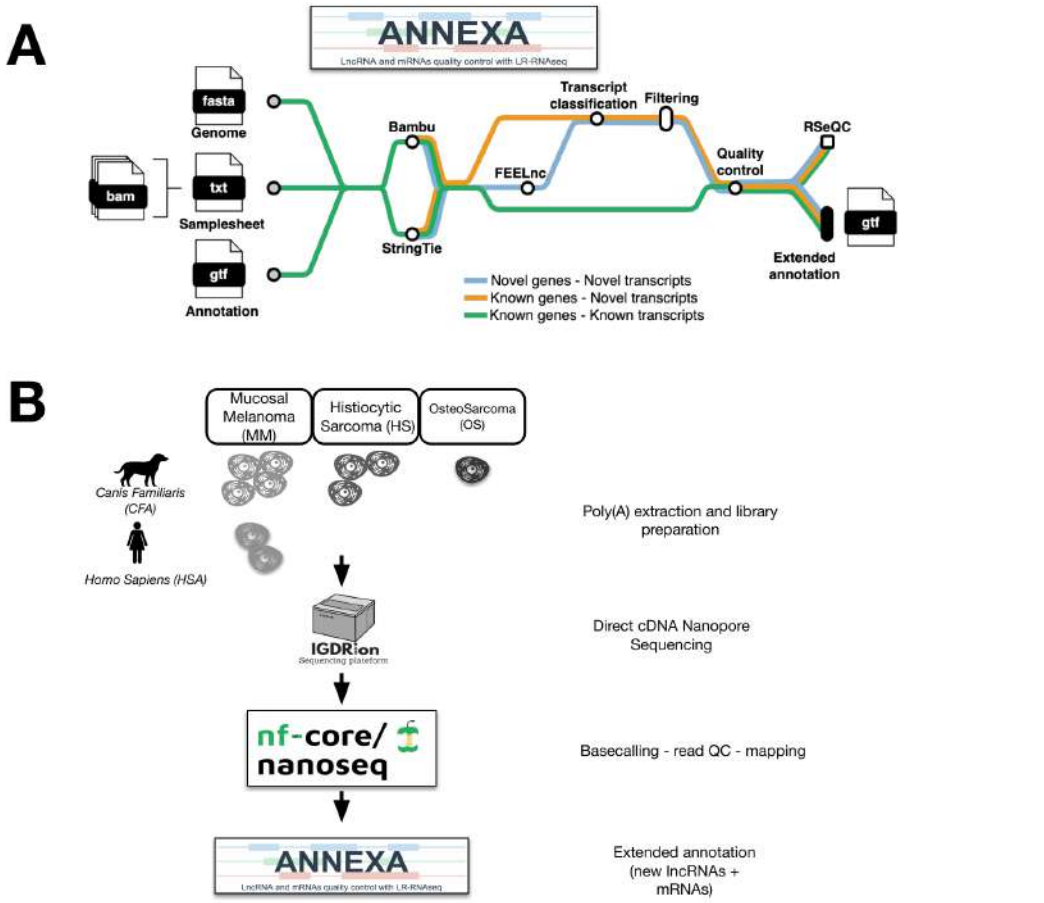


Fig. 1. A- Metromap of ANNEXA. **B**- Experimental design with species-specific input cancer cell lines and experimental and computational analyses.

Transcriptome reconstruction As recently demonstrated by the LR-GASP consortium, the choice of the bioinformatic tool to model known and novel genes/transcripts depends on the biological context, and more particularly, on the expected levels of annotation precision (controlling the number of novel false-positive transcripts) and sensitivity/recall (controlling the number of novel false-negative transcripts). Ideally, users should be able to select the appropriate tool based on the expected number of novel transcripts and the quality or completeness of the reference genome and annotation. To support this, ANNEXA enables the reconstruction and quantification of both known and novel transcripts using two distinct tools: Bambu [6] and StringTie2 [7]. In ANNEXA, several options are available for Bambu, including the ability to adjust the Novel Discovery Rate (NDR) threshold, a key metric that balances sensitivity and precision [6] across multiple samples. For instance, users can enhance sensitivity by increasing the default recommended NDR threshold with the `--bambu_threshold` option,

and/or include single-exon transcripts with the `--bambu_singleexon` option. Alternatively, users can choose to use the NDR threshold recommended by Bambu by enabling the `--bambu_rec_ndr` option, which replicates Bambu's default behaviour.
For StringTie2, ANNEXA uses the long reads assembly mode (`-L`) for the reconstruction step and the (`--merge`) option to produces a unified transcript set, while gene and transcript quantification (raw counts) are extracted using the `extractGeneExpression` function from the IsoformSwitchAnalyzeR program [11].
At the end of this module, ANNEXA integrates all genes and transcripts from the reference annotation along with the novel transcript models from StringTie2 or Bambu into an unfiltered annotation, referred to as *Extended_annotations.full.gtf*.

Coding potential and transcript classification Among the newly assembled transcripts, it is essential to annotate different RNA categories, particularly distinguishing protein-coding transcripts (mRNAs) from long non-coding RNAs (lncRNAs). To achieve this, ANNEXA integrates the FEELnc program [9] to predict the coding potential of all novel transcripts from novel genes (blue line, **Fig. 1.A**). To accelerate the FEELnc process, we previously demonstrated that training it with a subset of known mRNAs and lncRNAs from the reference annotation maintains high predictive performance [9], therefore allowing ANNEXA to integrate FEELnc with 3,000 reference lncRNAs and mRNAs.
To assess the potential impact of genomic alterations (mutations) on newly identified protein-coding transcripts, ANNEXA employs TransDecoder (https://github.com/TransDecoder/TransDecoder) with the `--single_best_only` option, ensuring that only the single best ORF per transcript is retained while reporting the corresponding CDS (Coding Determining Sequence) information in the final GTF file. While FEELnc is used to predict the coding potential and assign biotypes to all novel transcripts from novel genes (blue line in the metromap), TransDecoder is applied to both transcripts classified as coding by FEELnc (i.e., blue line) and for novel isoforms of known protein-coding genes (orange line in the metromap). Additionally, to classify novel transcripts with respect to the input reference annotation, ANNEXA uses Gffcompare [12], incorporating this information under the `class_code` attribute in the extended GTF. This enables users to efficiently extract specific transcript classes, such as `class k` corresponding to alternative isoforms extending known genes at the 5' or 3' ends or `class x`, corresponding to exonic antisense transcripts (often classified as antisense lncRNA biotype).

Transcript filtering and full-length assessment While long-read RNA sequencing (LR-RNAseq) has significantly improved transcriptome reconstruction by capturing full-length isoforms, it still exhibits biases that lead to fragmented transcript models, particularly at the 5' end/Transcription Start Sites (TSSs) [13].
To filter the *Extended_annotations.full.gtf* and remove novel incomplete transcripts, ANNEXA implements two filtering strategies based on (i) the Novel Discovery Rate (NDR) cut-off from the Bambu tool and (ii) the transforKmer cut-off, applicable to both Bambu and StringTie-derived transcripts. The transforKmer cut-off in ANNEXA evaluates the likelihood of all novel transcript TSSs using a species specific deep learning model pre-trained with DNABERT [14] and fine-tuned on a classification task using labelled TSSs from the reference annotation [15].
To adjust the stringency of the filtering process, users can choose between two modes: (i) the union

of the two filters (filtering `operation = union`), which retains transcripts passing at least one filter, or (ii) the intersection (filtering `operation = intersection`), which retains only transcripts passing both filters (See wiki ANNEXA).
At the end of this module, ANNEXA outputs an additional more stringent annotation, referred to as *Extended_annotations.filter.gtf*.

Quality Control The two resulting annotation files (*full* and *filtered*) may contain thousands of novel isoforms, including both mRNAs and lncRNAs, which require thorough inspection for final quality control (QC). Inspired by the SQANTI tool [16], ANNEXA computes multiple features by comparing known (*i.e.*, matching the reference annotation) and novel (*i.e.*, reconstructed by Bambu or StringTie) genes, transcripts, and exons, collectively referred to as annotated elements (AE) (See ANNEXA Wiki).

These features can be broadly categorized into two groups:

- Structural metrics of AEs: including the number of known and novel AEs, AE length distribution, proportion of single- versus multi-isoform genes (at the gene level), and proportion of single- versus multi-exonic transcripts (at the transcript level).
- Quantification-related metrics: such as the distribution of gene counts across input samples or the breadth of expression of the AEs.

Additionally, an optional QC feature assesses gene body coverage by sample reads using the RSeQC pipeline [17].
All QC indicators are generated as CSV files, which serve as input for ANNEXA to compile and visualise the data in a comprehensive final QC report (for illustration in **Fig.2**).

ANNEXA implementation ANNEXA is implemented in Nextflow [18] and integrates scripts written in multiple languages, including R, Python and Bash. The Nextflow framework is designed to facilitate the development of reproducible, scalable, and portable analysis workflows. It is available as both Docker and Singularity containers and has been successfully executed on standalone computers and high-performance computing clusters using SGE or SLURM. The project is open-source, with its code accessible on GitHub : https://github.com/IGDRion/ANNEXA/.

Long-Read Sequencing protocol of cancer cell lines

Experimental methods Ten LR-RNAseq experiments were conducted (**Fig.1.B**) with eight canine cancer cell lines from three canine cancer types: Mucosal Melanoma (MM, n=4), Histiocytic Sarcoma (HS, n=3) and Osteosarcoma (OS, n=1) and two human cancer cell lines also originating from MM patients. For all samples, RNA was extracted from ∼15 million cells using the NucleoSpin RNA kit (Macherey-Nagel). Library preparation was performed according to manufacturer's protocol (Oxford Nanopore Technologies, ONT) with the direct cDNA Sequencing Kit (SQK-DCS109). Sequencing was done using MinION Flow Cells (FLO-MIN106D) with GridION device from the IGDRION platform (https://igdr.univ-rennes.fr/igdrion).

Computational methods Basecalling of `fast5` files was done with guppy (version 6.0.0) and the `nf-core/nanoseq` pipeline (version 3.1.0) from the nf-core community [19] was used to do all primary bioinformatic analyses. Briefly, this included the quality control (QC) of the reads with the

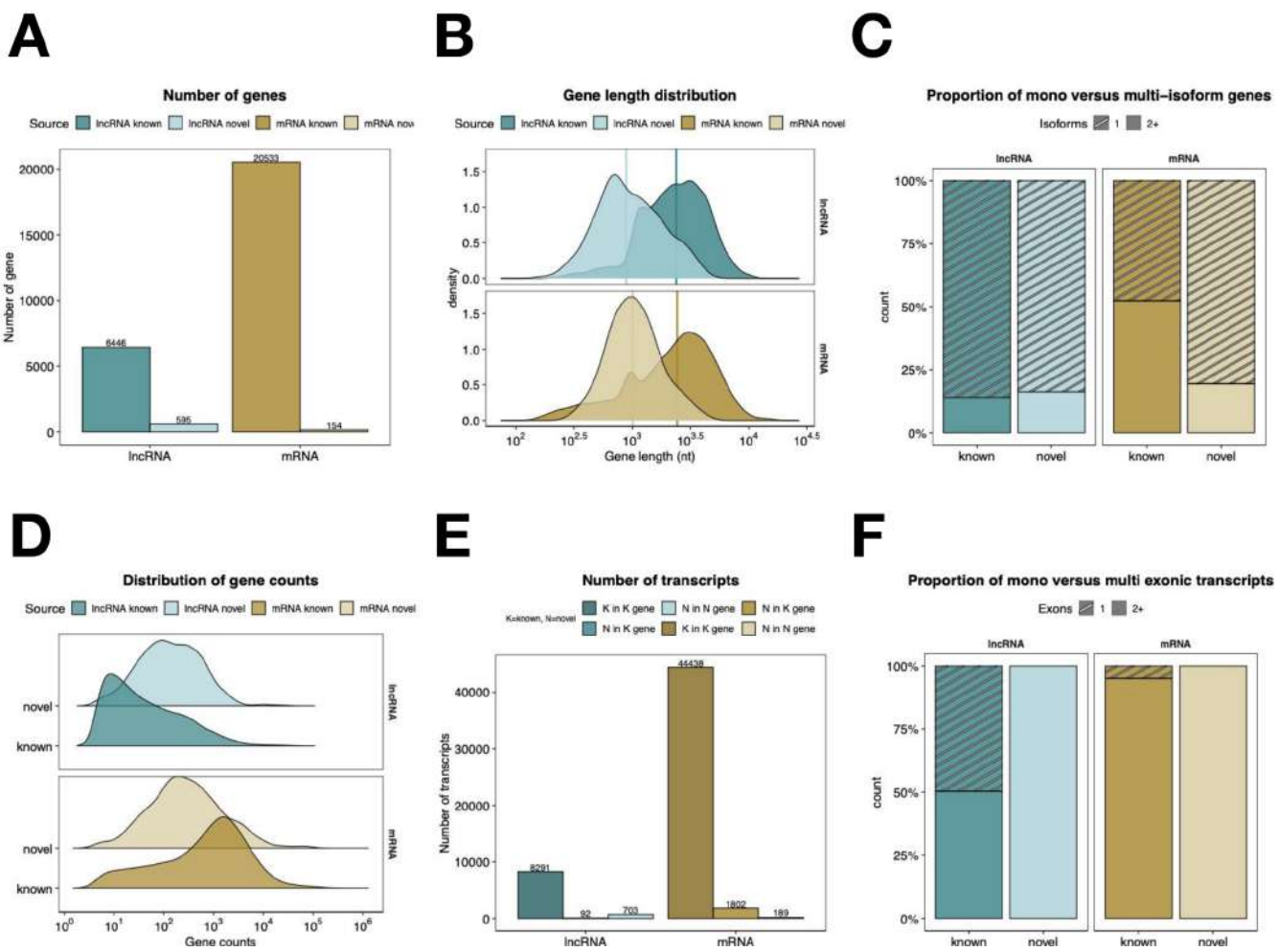


Fig. 2. Automatic QC report from ANNEXA. **A**- Number of known and novel genes. **B**- Gene length distribution. **C**- Proportion of gene mono- versus multi-isoform(s). **D**- Distribution of gene counts. **E**- Number of transcripts in known (K) and novel (N) genes. **F**- Proportion of transcript mono/single- versus multi-exonic.

nanoplot (version 1.41.6) [20] and multiqc (version 1.9) [21] programs and the mapping of the `fastq` files onto human and canine genomes using minimap2 software (version 2.15-r905) [22]. For the ten samples, we considered all reads having a Qscore >5 (**Table1**). Intersection between all genomics features (i.e. TSS and CAGE data) was done using bedtools (version 2.27.1).

Sample	Total Reads	Median Length	Primary Mapped	% Primary Mapped
CFA-MM1	4,002,891	806	3,408,273	85.15
CFA-MM2	8,843,341	769	6,253,908	70.72
CFA-MM3	7,108,351	943	5,901,302	83.02
CFA-MM4	6,877,092	896	5,696,283	82.83
CFA-HS1	11,177,946	826	9,157,969	81.93
CFA-HS2	4,055,709	878	3,355,138	82.73
CFA-HS3	3,908,392	793	3,040,788	77.80
CFA-OS1	6,816,094	1,050	5,707,582	83.74
HSA-MM1	5,457,058	886	4,865,343	89.16
HSA-MM2	4,734,796	852	4,025,056	85.01

Tab. 1. Number of reads sequenced and aligned on canFam4 (CFA) and GRCh38 (HSA).

Benchmark analyses

Reference annotations and genomes Depending on the species, we compiled reference genomes and annotations downloaded from Ensembl [23], Refseq [24], Gencode [25] or original papers [26] [27] (**Table 2**).

Species	Assembly	Reference Annotation	Genes	Transcripts
Dog	canFam3	Ensembl	30,951	60,994
Dog	canFam4	Ensembl	30,653	56,403
Dog	canFam4	UU (University Uppsala)	29,535	158,561
Dog	canFam4	Refseq	43,427	103,619
Dog	canFam5	Ensembl	26,037	48,737
Dog	canFam6	Ensembl	29,992	53,113
Human	GRCh38	Gencode	86,402	412,034
Human	GRCh38	CHESS	63,755	168,451

Tab. 2. Species, Assembly, Reference Annotation with Gene and Transcript Information

Transcriptome reconstruction benchmarking ANNEXA was run using Bambu and StringTie on canine samples with the canFam4 genome assembly and the Ensembl annotation, and on the human samples with the GRCh38 assembly and CHESS or Gencode annotations. For the Bambu runs, we discarded single exon transcripts and used the recommended NDR of the tool. The full and filtered extended annotations produced after the runs were filtered to only keep genes and transcripts expressed in at least one of each sample (raw count $>$0) and were compared to the initial reference annotations using the gffcompare tool version v0.12.6 ([12]) which provides sensitivity and precision metrics at the exon, transcript and locus level. At each level, reconstructed elements matching the reference annotation were considered true positives (TP), while those absent from the reference were classified as false positives (FP). Conversely, features present in the reference but missing from the extended annotations were considered as false negatives (FN). We thus defined Precision as TP /(TP + FP) and Recall or Sensitivity as TP/(TP+FN).

Orthology analysis Canine and human novel genes and transcripts were mapped on target genomes (GRCh38 and CanFam4, respectively) using the liftoff program [28] with default parameters. Then, query elements were classified into three classes with respect to the target reference annotation (Gencode and Ensembl, respectively): unmapped (query gene not mapped to target genome), mapped_unknownGenes (query gene mapping to intergenic regions) and mapped_knownGenes (query gene mapping to known genes from target annotation *e.g.* Gencode for query dog gene and Ensembl for human query gene).

Results

In a comparative oncology study, we produced LR-RNASeq data from both human (n=2) and canine (n=8) cancer cell lines (**Fig.1.B**) and applied ANNEXA using different genome assemblies and reference annotations for each species in order to test the robustness of the pipeline in its ability to extend and to quality control these annotations.

Effect of species-specific reference transcriptome for the identification of known and novel genes/transcripts Analysis of known and novel genomic features (genes and transcripts) recon-

structed by Bambu and StringTie (with their default parameters) revealed striking differences with respect to the tool used and the completeness of the input reference annotation (**Fig.3**).

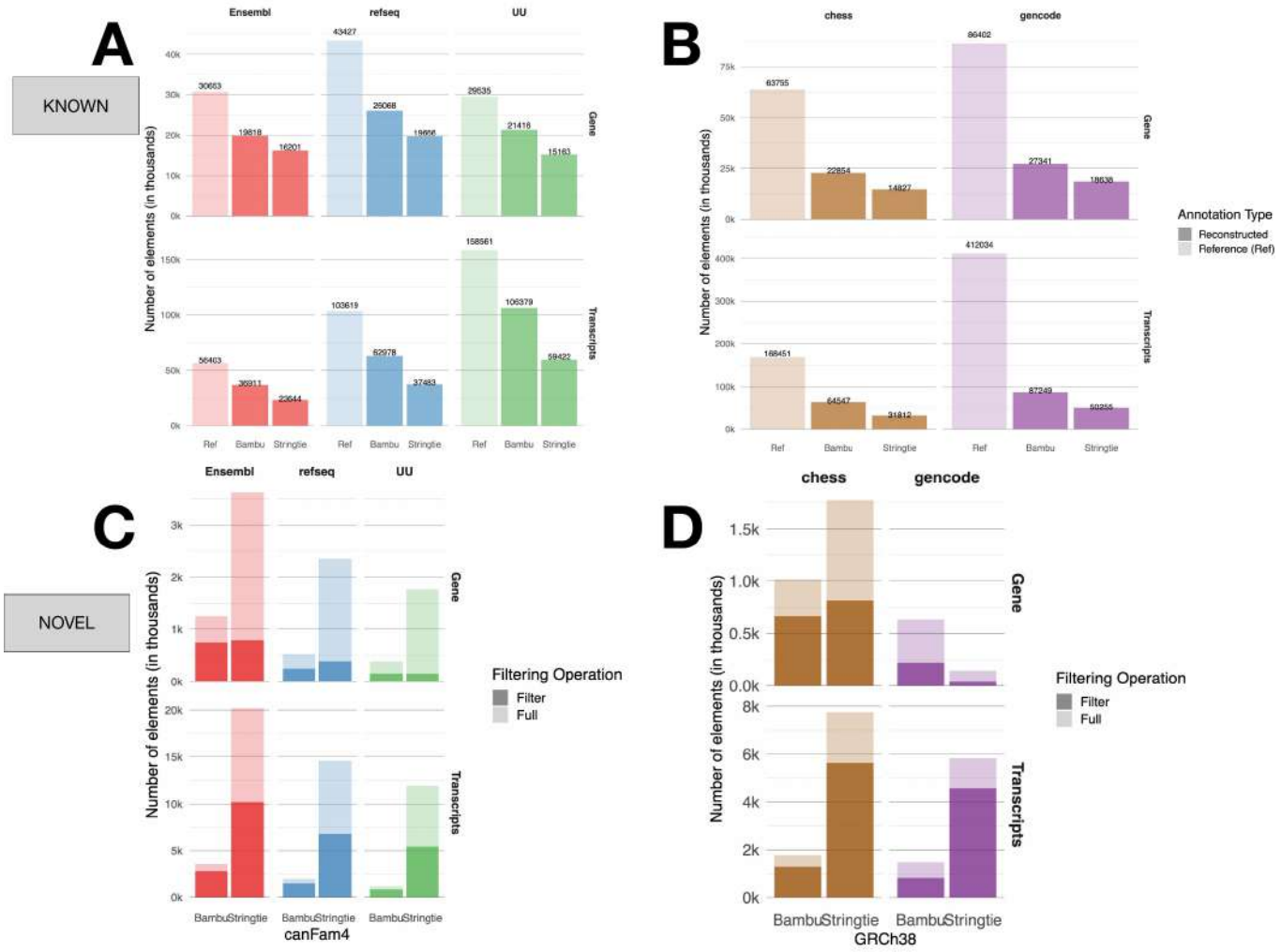


Fig. 3. Benchmark analysis between Bambu and StringTie for dog (left panels) and human (right panels) for known (top panel) and novel (bottom panel) genes/transcripts. For known elements (top), number of genes/transcripts from the reference annotation (light bar) are separated from reconstructed genes/transcripts (dark bar). **A**- Number of known canine genes/transcripts retrieved from three reference annotations (Ensembl - red, refseq - blue and UnivUppsala - green). **B**- Number of known human genes/transcripts retrieved from two reference annotations (Chess - brown and Gencode - purple). **C**- Number of novel dog genes/transcripts. Full set of elements (light bar) are separated from ANNEXA's filtered (intersection) set (dark bar). **D**- Number of novel human genes/transcripts. Full set of elements (light bar) are separated from ANNEXA's filtered (intersection) set (dark bar).

Known genes and transcripts For the canine genome (canFam4), Bambu reconstructed more known elements than StringTie, consistently across all three reference annotations tested (Ensembl, Refseq, UU), with this difference being the most noticeable for the University of Uppsala (UU) annotation (21,416 genes and 106,378 transcripts reconstructed with Bambu, 15,163 and 59,422 with StringTie) (**Fig.3.A**). For the human genome (GRCh38), Bambu also identified more known elements than StringTie across the two annotations (CHESS and Gencode) with this difference being most pronounced for Gencode (27,341 genes and 87,249 transcripts reconstructed with Bambu compared to

18638 and 50,255 respectively with StringTie) (**Fig.3.B**). Compared to dogs, the relatively low percentage of features reconstructed from a reference annotation in human (for example, only 21% of Gencode transcripts are reconstructed by Bambu) could be explained by the fact that human reference annotations are more complete than in dogs and also because our experimental design contained more LR-RNAseq samples in dogs (n=8) versus human (n=2).

Novel genes and transcripts For the canine genome (canFam4), StringTie consistently identified more novel elements than Bambu, independently of the three reference annotations tested (Ensembl, RefSeq and UU). However, we observed that this difference is most pronounced using the Ensembl annotation, where Bambu detected 3,134 novel genes and 3,600 novel transcripts compared to the StringTie set of genes and transcripts (12,671 and 20,184, respectively). Using the UU canine annotation yielded the lowest number of novel genes and transcripts identified by both StringTie (7,791 and 11,900) and Bambu (1,070 and 1,150), likely due to the greater initial isoform completeness of the UU reference annotation compared to Ensembl and RefSeq, thereby reducing the number of newly detected transcripts (**Fig.3.C**). In the human genome (GRCh38), using the Gencode annotation as reference produced substantially fewer novel elements than with CHESS, as expected given the higher number of genes and transcripts cataloged in the latest version of Gencode (**Fig.3.D**). Yet, at the gene level, we observed an intriguing pattern where StringTie annotates more novel genes with CHESS compared to Gencode, whereas the opposite trend occurs with Bambu ($\sim$1000 novel genes for Bambu and 1,800 for StringTie with the CHESS annotation as compared to 631 and 145 with Gencode, respectively). Regarding ANNEXA filtering operations, which evaluated the full-lengthness of reconstructed genomic elements (see **Methods**), we observed that it dramatically reduced the number of novel elements identified by both tools. On average in dogs, 50% of Bambu and 16% of StringTie novel genes are being conserved after ANNEXA's check for TSS validity. The proportion of StringTie's filtered transcripts is more pronounced with Ensembl (dog) and CHESS (human) annotations, suggesting that these novel transcripts may contain a higher proportion of incomplete transcripts. Together, these findings highlight that reference annotation selection significantly impacts novel element discovery, with StringTie demonstrating particular sensitivity to this choice, especially in less well-annotated genomes such as canFam4. In addition, this shows that Bambu consistently demonstrated stronger performance in terms of precision and recall, particularly after applying ANNEXA's filtering steps for gene and transcript completeness (**Supp. Fig. 1**)

Effect of species-specific reference transcriptome for the identification of known mRNAs and lncRNAs Compared to protein-coding genes (mRNAs), long non-coding RNAs (lncRNAs) are considered particularly challenging to annotate due to their low expression levels and tissue specificity [29] [30]. As in our previous analyses, we evaluated the ability of Bambu and StringTie to reconstruct known genes and transcripts starting from long-read RNA sequencing (LR-RNASeq) data, this time distinguishing lncRNAs from mRNAs (**Fig.4**).

In dogs (canFam4, Ensembl annotation), we showed that Bambu reconstructed a greater number of mRNAs and lncRNAs from the reference annotation than StringTie, with 13,239 vs. 9,930 genes for mRNAs and 3,314 vs. 2,525 for lncRNAs (**Fig.4.A**). At the transcript level, we also noticed that the difference between Bambu and StringTie in term of reconstructed transcripts is more pronounced

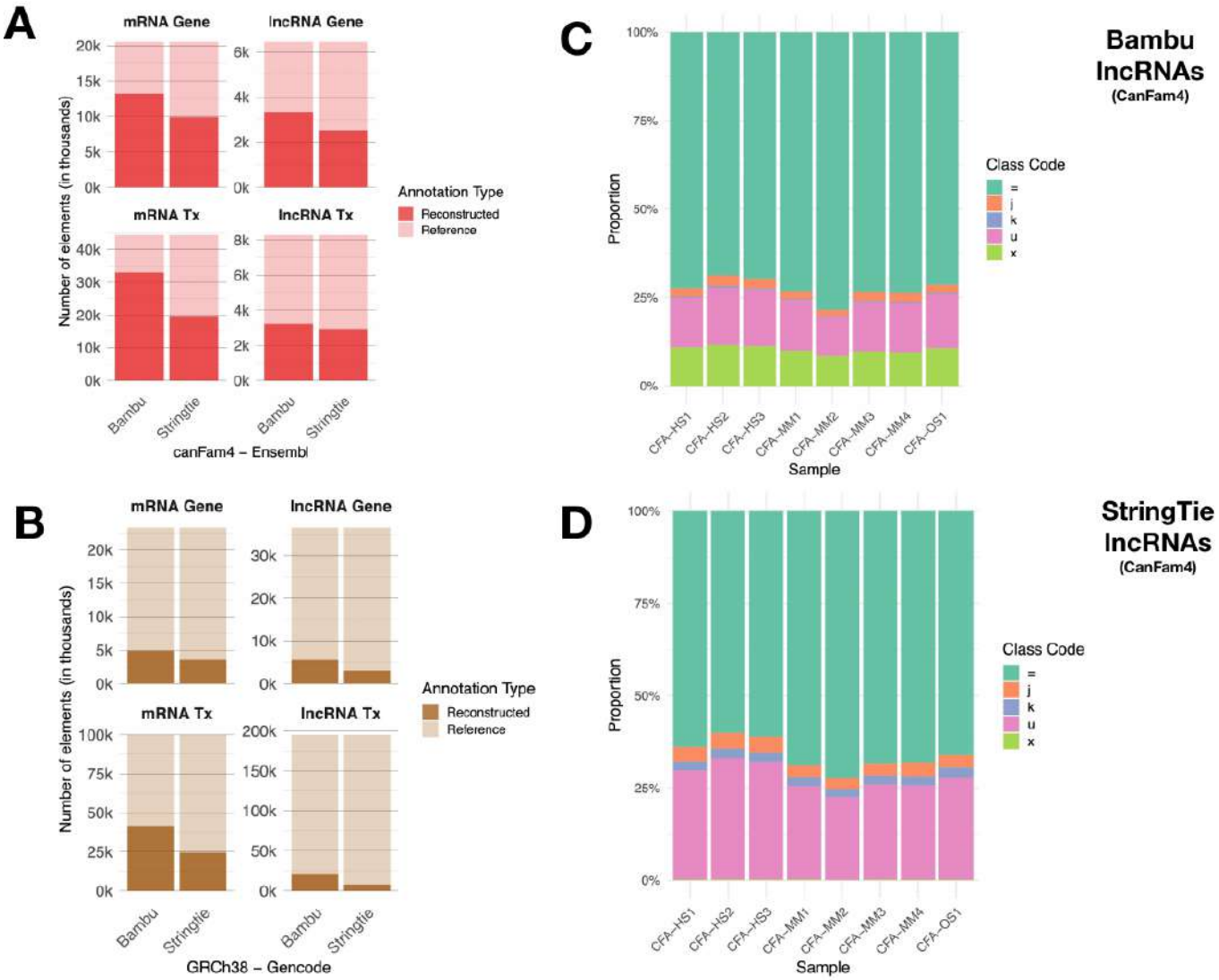


Fig. 4. Annotation of lncRNAs/mRNAs by Bambu and StringTie. **A**- Number of known canine genes (top) and transcripts (bottom) for mRNA (left) and lncRNAs (right) annotated by Ensembl. **B**- Number of known human genes (top) and transcripts (bottom) for mRNA (left) and lncRNAs (right) annotated by Gencode. **C**- Class code distribution of canine lncRNAs annotated by Bambu. **D**- Class code distribution of canine lncRNAs annotated by StringTie. For panel C and D, class_codes = (green) : transcript isoforms matching the reference annotation, j (orange): novel spliced isoforms, k (purple): extension of reference gene, u (pink): intergenic transcripts and x (light green): antisense transcripts.

for mRNAs than for lncRNAs biotypes. A similar trend was observed in humans (GRCh38, Gencode annotation), where Bambu identified more genes and transcripts from the latest Gencode annotation than StringTie (**Fig.4.B**). However, the proportion of reconstructed known genes was lower in humans than in dogs, again likely due to the smaller number of human LR-RNASeq data available as input.

Using ANNEXA's transcript classification module, we then categorized both known (`class_code =`) and novel (all other `class_code`) mRNAs and lncRNAs annotated by the two tools from the eight canine and two human LR-RNASeq samples. Interestingly, this analysis revealed distinct patterns in the classification of novel canine lncRNAs: Bambu predominantly assigned them to `class_code u` (intergenic lncRNAs or lincRNAs, 14%) and `class_code x` (antisense lncRNAs, 10%), whereas StringTie

primarily annotated novel lncRNAs as lincRNA (27%) and extension of known lncRNAs (2.4%), with very few classified as antisense (n=4) in dogs (**Fig.4.C** and **Fig.4.D**).

Biological application of ANNEXA Over the past decade, canine models have gained recognition as valuable spontaneous and immunocompetent systems for studying human cancers, particularly histiocytic sarcomas (HS) [31] and mucosal melanoma (MM) [32]. For MM, although rare in humans, it represents the most prevalent oral malignancy in dogs and exhibits notable clinical, biological, and genetic parallels with its human counterpart [33]. Leveraging ANNEXA's ability to balance precision and recall, we applied it to both canine and human data. In dogs, using a relaxed Bambu NDR threshold (NDR = 1) allowed to identify 9,612 novel genes (8,713 lncRNAs and 899 mRNAs) across the eight long-read RNA-Seq datasets (see **Methods**). Applying ANNEXA's TSS validity filter significantly reduced the number of novel genes to 749 (595 lncRNAs and 154 mRNAs), indicating a likely high rate of false positives in the unfiltered set. Notably, the proportion of gene TSSs validated by orthogonal datasets, such as CAGE (Cap Analysis of Gene Expression) from the DogA consortium [34], was significantly higher in the filtered set (52%, with 311 novel lncRNAs and 75 mRNAs validated) compared to the full dataset (9.3%, with 903 novel lncRNAs and 186 mRNAs validated). We also sought to assess whether novel genes in both humans and dogs could be conserved through evolution as a proxy for functional evidence [35]. We thus mapped the 9,614 novel canine genes (Ensembl - canFam4) on the human genome assembly used for the extended human annotation from Gencode and found 3,709 (38.6%) that could be mapped to GRCh38. Among these, 3,268 (88%) correspond to human genes from the Gencode extended reference annotation and thus represent novel orthologous relationships between novel dog and known human genes. Notably, we identified five novel canine lncRNA genes that also map to novel human genes (**Supp. Fig. 2**), with two of them being classified as protein coding in human. Although these genes were not supported by CAGE data, they exhibited relatively high expression levels in both species, with mean read counts of 90.3 in dog and 66.1 in human MM samples. These conserved and expressed genes across species provide novel candidates for future functional validation and underscore the value of cross-species annotation to uncover novel, potentially functional elements.

Discussion

In this work, we introduce ANNEXA, an all-in-one tool that not only performs transcriptome reconstruction but also simultaneously ensures quality control of extended annotations. In addition, by uniquely integrating the comparative characterization of lncRNAs annotated from long-read RNA-seq data, ANNEXA enhances the profiling of these non-coding elements and refines their potential biological relevance. The comparison of two main transcriptome discovery tools and quantification methods reveals distinct trends depending on the reference annotation used. Our results show that Bambu consistently reconstructed more known genes and transcripts than StringTie, regardless of the reference annotation or the species (dog and human) analysed in this study. It also confirmed that the number of newly detected genomic elements is strongly influenced by the structure and coverage of the reference annotations. These findings align with the Long-read RNA-Seq Genome Annotation Assessment Project (LRGASP) Consortium's observations, which demonstrated that tools based on reference genomes perform optimally in well-annotated genomes [5]. The impact of reference annotation choice is also evident in the analysis of our human long-read RNASeq data. As

expected, using the latest Gencode release as reference produced fewer novel elements due to the higher number of genes and transcripts (86,402 and 412,034, respectively) already catalogued in its latest version. This underscores the importance of a comprehensive curated reference annotation to limit the detection of potentially artifact-driven elements.

Using ANNEXA's quality control module for known and novel long non-coding RNAs, we also observed this consistent trend where Bambu reconstructed a greater number of non-coding genes and transcripts than StringTie, both in dogs (canFam4, Ensembl) and humans (GRCh38, Gencode). However, the proportion of reconstructed known genes was lower in humans than in dogs for both tools, likely due to the more limited availability of human long-read RNA sequencing data in our study design, combined with the higher completeness of human reference annotations [8].

Additionally, the classification of novel lncRNAs revealed distinct patterns between the two tools. Bambu primarily categorised novel lncRNAs as intergenic (lincRNAs) or antisense, whereas StringTie assigned a higher proportion to lincRNAs and novel isoforms of known lncRNAs, with only a few lncRNAs classified as antisense. These differences in classification outputs underscore the impact of transcriptome reconstruction tools and suggest that integrating multiple approaches may provide a more comprehensive representation of lncRNA diversity. Consequently, the choice of reconstruction tool should be guided by specific research objectives. For example, researchers focusing on natural antisense transcripts (NATs) may benefit more from Bambu's capabilities, whereas those aiming to extend known lncRNA annotations might prefer StringTie.

Our analyses further emphasizes the need to carefully evaluate novel annotations generated by transcriptome discovery tools. The increase in detected genes and transcripts should be considered within the context of annotation robustness and biological validation. The LRGASP Consortium recommends incorporating additional orthogonal data and replicate samples when aiming to detect rare and novel transcripts. In our comparative oncology project, we used CAGE data from both dog and human samples to validate the full-length nature of novel transcripts. We found that only less than 10% of unfiltered novel transcripts (both coding and noncoding) were validated by at least one CAGE signal. Several factors could explain this relatively low proportion of TSS validation, including the fact that the CAGE sample conditions did not match those of our LR-RNASeq cancer cell lines and/or the earlier version of the Nanopore kit for library preparation (DCS109) could have led to truncated reads and, consequently, incomplete transcripts [4]. However, the use of ANNEXA's module for TSS validation increased the proportion of validated genes to over 50%, highlighting the importance of combining complementary experimental and computational approaches to refine annotations.

In summary, this study highlights ANNEXA's modularity in adjusting precision and sensitivity for accurate, context-specific coding and non-coding annotation. This flexibility is essential for subsequent experimental validation and for advancing our understanding of the role of the non-coding genome in biological processes.

Availability and implementation

ANNEXA is written in Nextflow DSL2 [18]. ANNEXA can run each process in conda environments as well as Docker or Apptainer containers, ensuring reproducibility and ease of use on different machines. The pipeline is available at https://github.com/IGDRion/ANNEXA. Input files and code to re-

produce figures of this paper are also available here :
https://github.com/IGDRion/ANNEXA/tree/main/Paper_Figures

Acknowledgements

The authors thank the Genouest bioinformatic core facility for providing us with the necessary storage and CPU/GPU resources to perform the analysis, the IGDRIon facility for long-read sequencing, members of the BIS group (https://igdr.univ-rennes.fr/biologie-silico-0) for useful discussions.

Funding Information

This study was supported by IGDR, University of Rennes, Region Bretagne with financial support from ITMO Cancer of Aviesan within the framework of the 2021-2030 Cancer Control Strategy, and funds from Inserm (C20013NS), Cancéropole Grand Ouest (CGO), Ligue Regionale contre le cancer du Grand Ouest and MSDAvenir foundation.

References

[1] Nurk S, Koren S, Rhie A, Rautiainen M, Bzikadze AV, Mikheenko A, et al. The complete sequence of a human genome. Science (New York, NY). 2022 Apr;376(6588):44-53.

[2] Steijger T, Abril JF, Engström PG, Kokocinski F, Hubbard TJ, Guigó R, et al. Assessment of transcript reconstruction methods for RNA-seq. Nature Methods. 2013;10(12):1177-84. Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3851240/.

[3] Conesa A, Madrigal P, Tarazona S, Gomez-Cabrero D, Cervera A, McPherson A, et al. A survey of best practices for RNA-seq data analysis. Genome Biology. 2016;17:13. Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4728800/.

[4] Sessegolo C, Cruaud C, Da Silva C, Cologne A, Dubarry M, Derrien T, et al. Transcriptome profiling of mouse samples using nanopore sequencing of cDNA and RNA molecules. Scientific Reports. 2019 Oct;9(1):14908. Publisher: Nature Publishing Group. Available from: https://www.nature.com/articles/s41598-019-51470-9.

[5] Pardo-Palacios FJ, Wang D, Reese F, Diekhans M, Carbonell-Sala S, Williams B, et al. Systematic assessment of long-read RNA-seq methods for transcript identification and quantification. Nature Methods. 2024 Jul;21(7):1349-63. Publisher: Nature Publishing Group. Available from: https://www.nature.com/articles/s41592-024-02298-3.

[6] Chen Y, Sim A, Wan YK, Yeo K, Lee JJX, Ling MH, et al. Context-aware transcript quantification from long-read RNA-seq data with Bambu. Nature Methods. 2023 Aug;20(8):1187-95. Publisher: Nature Publishing Group. Available from: https://www.nature.com/articles/s41592-023-01908-w.

[7] Kovaka S, Zimin AV, Pertea GM, Razaghi R, Salzberg SL, Pertea M. Transcriptome assembly from long-read RNA-seq alignments with StringTie2. Genome Biology. 2019 Dec;20(1):1-13. Number: 1 Publisher: BioMed Central. Available from: https://genomebiology.biomedcentral.com/articles/10.1186/s13059-019-1910-1.

[8] Kaur G, Perteghella T, Carbonell-Sala S, Gonzalez-Martinez J, Hunt T, Madry T, et al.. GENCODE: massively expanding the lncRNA catalog through capture long-read RNA sequencing. bioRxiv; 2024. Pages: 2024.10.29.620654 Section: New Results. Available from: https://www.biorxiv.org/content/10.1101/2024.10.29.620654v1.

[9] Wucher V, Legeai F, Hédan B, Rizk G, Lagoutte L, Leeb T, et al. FEELnc: a tool for long non-coding RNA annotation and its application to the dog transcriptome. Nucleic Acids Research. 2017 May;45(8):e57. Available from: https://doi.org/10.1093/nar/gkw1306.

[10] Guizard S, Miedzinska K, Smith J, Smith J, Kuo RI, Davey M, et al. nf-core/isoseq: simple gene and isoform annotation with PacBio Iso-Seq long-read sequencing. Bioinformatics. 2023 May;39(5):btad150. Available from: https://doi.org/10.1093/bioinformatics/btad150.

[11] Vitting-Seerup K, Sandelin A. IsoformSwitchAnalyzeR: analysis of changes in genome-wide patterns of alternative splicing and its functional consequences. Bioinformatics. 2019 Nov;35(21):4469-71. Available from: https://doi.org/10.1093/bioinformatics/btz247.

[12] Pertea G, Pertea M. GFF Utilities: GffRead and GffCompare. F1000Research; 2020. Available from: https://f1000research.com/articles/9-304.

[13] Tardaguila M, de la Fuente L, Marti C, Pereira C, Pardo-Palacios FJ, Del Risco H, et al. SQANTI: extensive characterization of long-read transcript sequences for quality control in full-length transcriptome identification and quantification. Genome Research. 2018 Mar;28(3):396-411.

[14] Ji Y, Zhou Z, Liu H, Davuluri RV. DNABERT: pre-trained Bidirectional Encoder Representations from Transformers model for DNA-language in genome. Bioinformatics. 2021 Feb;37(15):2112-20. Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC11025658/.

[15] Karollus A, Hingerl J, Gankin D, Grosshauser M, Klemon K, Gagneur J. Species-aware DNA language models capture regulatory elements and their evolution. Genome Biology. 2024 Dec;25(1):1-21. Number: 1 Publisher: BioMed Central. Available from: https://genomebiology.biomedcentral.com/articles/10.1186/s13059-024-03221-x.

[16] Pardo-Palacios FJ, Arzalluz-Luque A, Kondratova L, Salguero P, Mestre-Tomás J, Amorín R, et al. SQANTI3: curation of long-read transcriptomes for accurate identification of known and novel isoforms. Nature Methods. 2024 May;21(5):793-7. Publisher: Nature Publishing Group. Available from: https://www.nature.com/articles/s41592-024-02229-2.

[17] Wang L, Wang S, Li W. RSeQC: quality control of RNA-seq experiments. Bioinformatics. 2012 Aug;28(16):2184-5. Available from: https://doi.org/10.1093/bioinformatics/bts356.

[18] Di Tommaso P, Chatzou M, Floden EW, Barja PP, Palumbo E, Notredame C. Nextflow enables reproducible computational workflows. Nature Biotechnology. 2017 Apr;35(4):316-9. Publisher: Nature Publishing Group. Available from: https://www.nature.com/articles/nbt.3820.

[19] Ewels PA, Peltzer A, Fillinger S, Patel H, Alneberg J, Wilm A, et al. The nf-core framework for community-curated bioinformatics pipelines. Nature Biotechnology. 2020 Mar;38(3):276-8. Publisher: Nature Publishing Group. Available from: https://www.nature.com/articles/s41587-020-0439-x.

[20] De Coster W, Rademakers R. NanoPack2: population-scale evaluation of long-read sequencing data. Bioinformatics. 2023 May;39(5):btad311. Available from: https://doi.org/10.1093/bioinformatics/btad311.

[21] Ewels P, Magnusson M, Lundin S, Käller M. MultiQC: summarize analysis results for multiple tools and samples in a single report. Bioinformatics. 2016 Oct;32(19):3047-8. Available from: https://doi.org/10.1093/bioinformatics/btw354.

[22] Li H. Minimap2: pairwise alignment for nucleotide sequences. Bioinformatics. 2018 Sep;34(18):3094-100. Available from: https://doi.org/10.1093/bioinformatics/bty191.

[23] Dyer SC, Austine-Orimoloye O, Azov AG, Barba M, Barnes I, Barrera-Enriquez VP, et al. Ensembl 2025. Nucleic Acids Research. 2025 Jan;53(D1):D948-57. Available from: https://doi.org/10.1093/nar/gkae1071.

[24] O'Leary NA, Wright MW, Brister JR, Ciufo S, Haddad D, McVeigh R, et al. Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. Nucleic Acids Research. 2016 Jan;44(Database issue):D733-45. Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4702849/.

[25] Mudge JM, Carbonell-Sala S, Diekhans M, Martinez JG, Hunt T, Jungreis I, et al. GENCODE 2025: reference gene annotation for human and mouse. Nucleic Acids Research. 2025 Jan;53(D1):D966-75.

[26] Wang C, Wallerman O, Arendt ML, Sundström E, Karlsson Nordin J, et al. A novel canine reference genome resolves genomic architecture and uncovers transcript complexity. Communications Biology. 2021 Feb;4(1):1-11. Publisher: Nature Publishing Group. Available from: https://www.nature.com/articles/s42003-021-01698-x.

[27] Varabyou A, Sommer MJ, Erdogdu B, Shinder I, Minkin I, Chao KH, et al. CHESS 3: an improved, comprehensive catalog of human genes and transcripts based on large-scale expression data, phylogenetic analysis, and protein structure. Genome Biology. 2023 Oct;24(1):249. Available from: https://doi.org/10.1186/s13059-023-03088-4.

[28] Shumate A, Salzberg SL. Liftoff: accurate mapping of gene annotations. Bioinformatics. 2021 Jul;37(12):1639-43. Available from: https://doi.org/10.1093/bioinformatics/btaa1016.

[29] Kainth AS, Haddad GA, Hall JM, Ruthenburg AJ. Merging short and stranded long reads improves transcript assembly. PLOS Computational Biology. 2023 Oct;19(10):e1011576. Publisher: Public Library of Science. Available from: https://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1011576.

[30] Derrien T, Johnson R, Bussotti G, Tanzer A, Djebali S, Tilgner H, et al. The GENCODE v7 catalog of human long noncoding RNAs: Analysis of their gene structure, evolution, and expression. Genome Research. 2012 Jan;22(9):1775-89. Company: Cold Spring Harbor Laboratory Press Distributor: Cold Spring Harbor Laboratory Press Institution: Cold Spring Harbor Laboratory Press Label: Cold Spring Harbor Laboratory Press Publisher: Cold Spring Harbor Lab. Available from: http://genome.cshlp.org/content/22/9/1775.

[31] Hédan B, Rault M, Abadie J, Ulvé R, Botherel N, Devauchelle P, et al. PTPN11 mutations in canine and human disseminated histiocytic sarcoma. International Journal of Cancer. 2020;147(6):1657-65. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/ijc.32991. Available from: https://onlinelibrary.wiley.com/doi/abs/10.1002/ijc.32991.

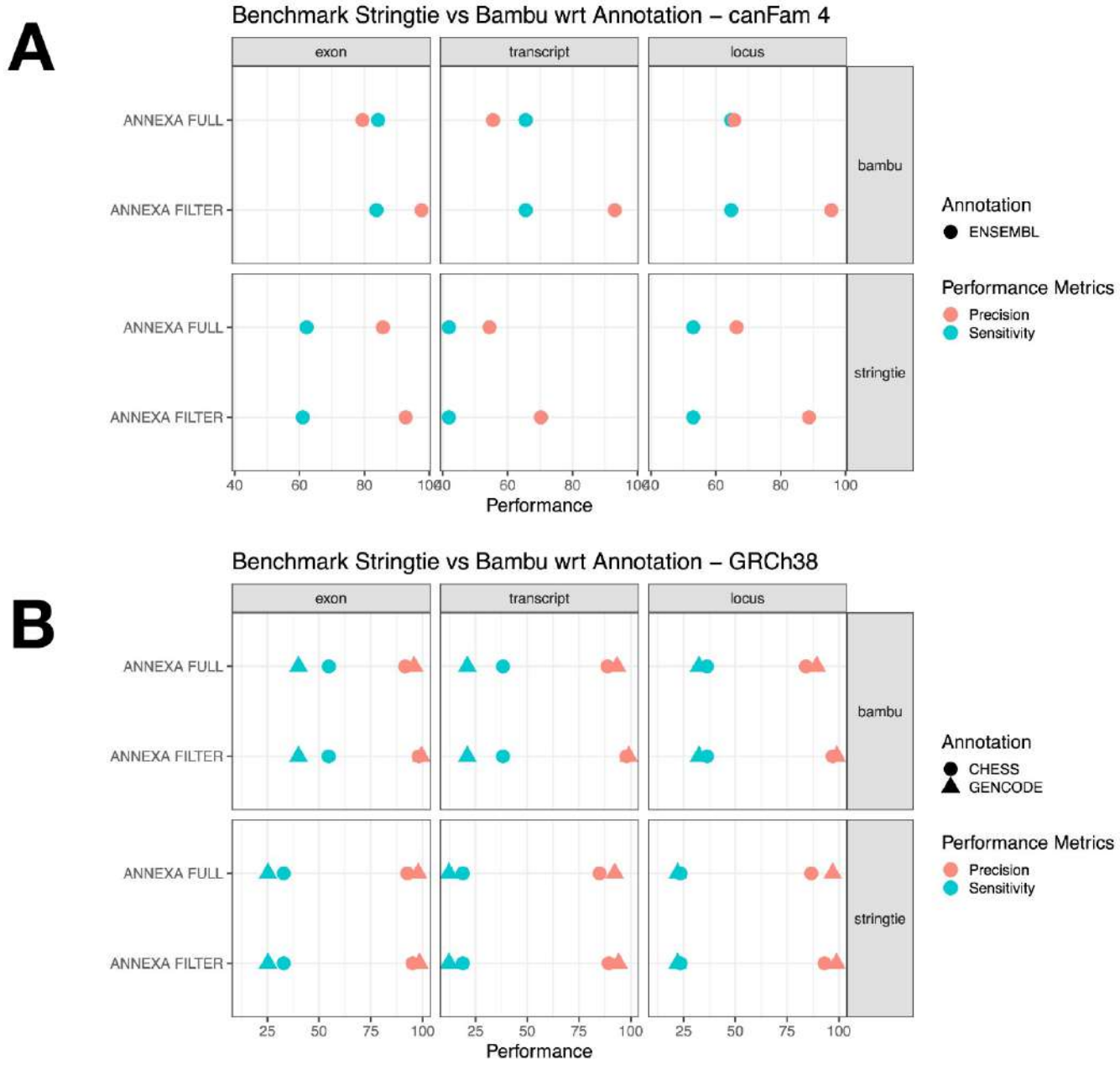
[32] Prouteau A, André C. Canine Melanomas as Models for Human Melanomas: Clinical, Histological, and Genetic Comparison. Genes. 2019 Jul;10(7):501. Number: 7 Publisher: Multidisciplinary Digital Publishing Institute. Available from: https://www.mdpi.com/2073-4425/10/7/501.

[33] Prouteau A, Mottier S, Primot A, Cadieu E, Bachelot L, Botherel N, et al. Canine Oral Melanoma Genomic and Transcriptomic Study Defines Two Molecular Subgroups with Different Therapeutical Targets. Cancers. 2022 Jan;14(2):276. Number: 2 Publisher: Multidisciplinary Digital Publishing Institute. Available from: https://www.mdpi.com/2072-6694/14/2/276.

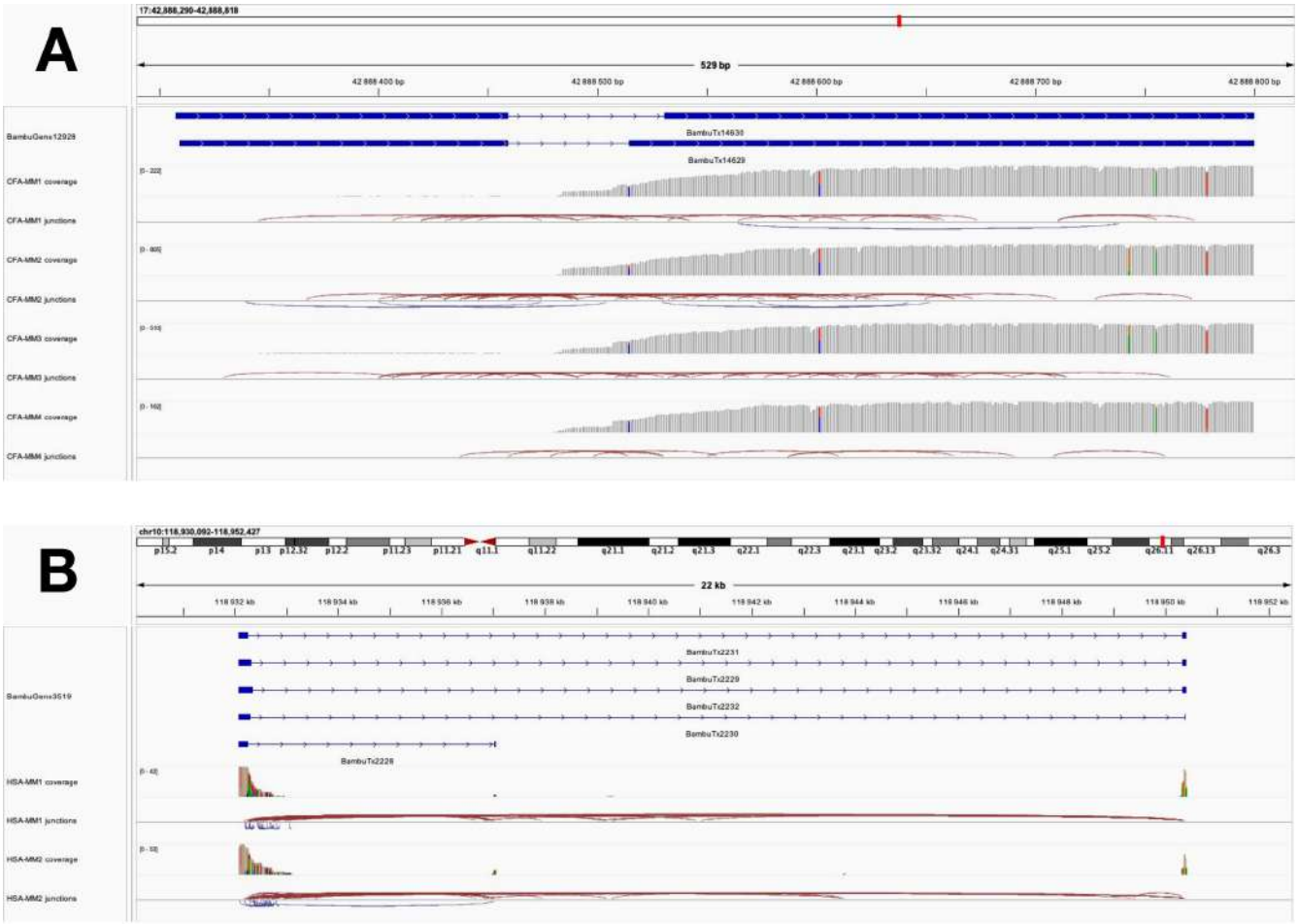
[34] Hörtenhuber M, Hytönen MK, Mukarram AK, Arumilli M, Araujo CL, Quintero I, et al. The DoGA consortium expression atlas of promoters and genes in 100 canine tissues. Nature Communications. 2024 Oct;15(1):9082. Publisher: Nature Publishing Group. Available from: https://www.nature.com/articles/s41467-024-52798-1.

[35] Ulitsky I. Evolution to the rescue: using comparative genomics to understand long non-coding RNAs. Nature Reviews Genetics. 2016 Oct;17(10):601-14. Publisher: Nature Publishing Group. Available from: https://www.nature.com/articles/nrg.2016.85.

Supplementary Figures



Supplementary Figure 1. Comparison of precision (red) and sensitivity (green) metrics when using ANNEXA with Bambu or Stringtie on the canine (A) or human (B) LR-RNAseq data.



Supplementary Figure 2. IGV screenshot of a novel orthologous gene both annotated in dog (A) and human (B). Long reads aligned from four mucosal melanoma (MM) samples in dogs (canFam4) and two MM samples in human (GRCh38) are represent below annotations.

SpecPeptidOMS Directly and Rapidly Aligns Mass Spectra on Whole Proteomes and Identifies Peptides That Are Not Necessarily Tryptic: Implications for Peptidomics

Émile BENOIST[1], Géraldine JEAN[1], Hélène ROGNIAUX[2], Guillaume FERTIN[1] and Dominique TESSIER[2]

[1] Nantes Université, CNRS, LS2N, UMR 6004, F-44000 Nantes, France

[2] INRAE, PROBE Research Infrastructure, BIBS Facility, F-44300 Nantes, France; INRAE, UR1268 Biopolymères Interactions Assemblages, F-44316 Nantes, France

Corresponding author: geraldine.jean@univ-nantes.fr

Reference paper: **Benoist et al. (2025) SpecPeptidOMS Directly and Rapidly Aligns Mass Spectra on Whole Proteomes and Identifies Peptides That Are Not Necessarily Tryptic: Implications for Peptidomics.** ***Journal of Proteome Research*****.** https://pubs.acs.org/doi/full/10.1021/acs.jproteome.4c00870

Keywords Peptidomics, Proteomics, Mass spectrometry, MS2 spectra, Dynamic programming.

Abstract *SpecPeptidOMS directly aligns peptide fragmentation spectra to whole and undigested protein sequences. The algorithm was specifically and initially designed for peptidomics, where the aim is to identify peptides that do not result from the hydrolysis of a known protein and therefore, whose termini cannot be predicted. Thus, SpecPeptidOMS can perform alignments starting and ending anywhere in the protein sequence. The underlying computational method of SpecPeptidOMS, which is based on a dynamic programming approach, was drastically optimized. As a result, SpecPeptidOMS can process around 12,000 spectra per hour on an ordinary laptop, with alignment performed against the entire human proteome. The performance of SpecPeptidOMS was first evaluated on a publicly available data set of (nontryptic) synthetic mass spectra. Accuracy was estimated by considering the results obtained by MaxQuant on the same data set as the "ground truth". A second series of tests on a larger, well-known proteomics data set (HEK293) highlighted SpecPeptidOMS' additional ability to search for open modifications, a feature of interest in peptidomics but also more broadly in conventional proteomics. SpecPeptidOMS is open-source, cross-platform (written in Java), and freely available.*

Highlight

Peptidomics is an emerging field of the omics sciences, that refers to the identification of the pool of peptides present in a biological fluid or tissue. The field has grown rapidly in the past few years because **endogenous peptides** have been established as **essential players in cellular processes** (e.g., signaling, immune response, intercellular communication, homeostasis, etc.) and potential biomarkers for a number of cellular disorders. **Nonendogenous peptides** are another group of circulating peptides that notably include food-derived peptides, and **food peptidomics** is arousing great interest. From a bioinformatics perspective, interpreting MS2 spectra in peptidomics adds complexity compared to proteomics because no assumption can be made about cleavage sites, meaning that **the peptide ends can lie at any amino acid of the protein**. Another characteristic of peptidomics is that **peptides can be highly modified**: this characteristic is almost inherent to their bioactivity, as modifications protect them from overly rapid proteolysis in biological fluids.

SpecPeptidOMS is a major upgrade of the SpecGlobX algorithm, which our group has recently proposed for the identification of multiple and unbiased modifications of tryptic peptides. Both algorithms fall in the family of the so-called **"open modification search" (OMS) methods** that have emerged over the past decade in proteomics. OMS methods have implemented **advanced computational optimization techniques** accelerating spectra comparison to a sufficient extent to widen the mass window of spectra comparison. As a result, those methods can identify peptides carrying unanticipated modifications. However, if the identification and localization of a single modification are successful, the presence of multiple modifications in a mass spectrum remains problematic. SpecGlobX, based on an **efficient dynamic programming algorithm**, can align pairs of spectra quickly while detecting **several modifications**. To limit execution time, SpecGlobX runs on a set of Peptide Spectrum Matches (PSMs) generated by another OMS search engine (SpecOMS, for example).

Compared to SpecGlobX, SpecPeptidOMS incorporates **two fundamental design changes** that deeply impact its capabilities. First, a **new and condensed representation of experimental spectra** allows direct alignment of MS2 spectra to undigested proteins without preconceptions about where the alignments should start and end in the protein. Second, a **drastic optimization** boosts the **execution time by several orders of magnitude** while keeping the memory requirements low. Importantly, SpecPeptidOMS retains the advantage inherited from SpecGlobX in **identifying peptides carrying multiple modifications** that had not been anticipated. However, while the quality of the SpecGlobX results were depending on the relevance of the set of PSMs used as input, **SpecPeptidOMS is independent of any other tool and evaluates all possible PSMs**. Then, the new design of SpecPeptidOMS has opened the way for efficient identification of MS2 spectra arising from peptidomics data sets.

The results obtained on two different data sets (one containing nontryptic peptides whose interpretation is approximately known, and the other corresponding to a large-scale proteomic data set already analyzed by several well-established software) are convincing of the **ability of SpecPeptidOMS to interpret spectra corresponding to peptides whose extremities are unknown, and possibly carrying modifications**. Thus, SpecPeptidOMS appears as a promising algorithm to interpret spectra in peptidomics, a field in which satisfactory tools are still lacking.

We believe this work to be relevant to the JOBIM community: first, because of the dedicated **computational design** of SpecPeptidOMS; second, because of its **ability to identify peptide sequences in the peptidomics context**. Our work combines strong methodological aspects, consideration of experimental conditions, and eagerness to accurately answer the biological problem at hand, which is to increase knowledge in the emerging peptidomics field.

We encourage the interested community to test SpecPeptidOMS: it is an **easy-to-use software** with few parameters to set up that does not require installation. With its command-line mode, SpecPeptidOMS is easy to integrate into a workflow.

Acknowledgments

This work was supported by the French National Agency for Research, ANR, through the ANR project PeptidOMS (ANR- 24-CE45-3296).

SVJedi-Tag : a novel method for genotyping large inversions with linked-read data

Mélody TEMPERVILLE[1], Fantine BENOIT[2], Claire MÉROT[2], Fabrice LEGEAI[1,3] and Claire LEMAITRE[1]

[1] Univ Rennes, Inria, CNRS, IRISA - UMR 6074, F-35000 Rennes, France
[2] Université de Rennes, CNRS, ECOBIO - UMR6553, 35000 Rennes, France
[3] IGEPP, INRAE, Institut Agro, University of Rennes, 35653 Le Rheu, France

Corresponding author: melody.temperville@inria.fr

Keywords Structural Variants, sequencing data analysis, variant genotyping, population genomics

Abstract *Structural Variants (SVs) are an important but overlooked aspect of genetic variation. In particular, inversions are known for their role in the evolution of biological diversity and particularly studied in non-model species using population data. One of the major steps in the study of SVs is genotyping. Linked-read data provide a cost-efficient alternative to long-reads to genotype many individuals, by combining the low sequencing cost of short reads with long-distance information thanks to the use of barcodes tagging long molecules. Whereas several methods have been proposed to discover SVs with linked-reads, there are currently no tool for genotyping with this type of sequencing data. In this paper, we present SVJedi-Tag, the first inversion genotyping method dedicated to linked-read data. We tested SVJedi-Tag on simulated and real linked-read data in the seaweed fly* Coelopa frigida*, and showed that SVJedi-Tag is able to genotype with high accuracy large inversions above 25 kb, with a read depth as low as 3X.*

Introduction

Accurately detecting and characterizing genetic variation is essential for all aspects of genomics including medical research, ecological and evolutionary genomics, application for food production, and conservation. While most studies have long focused on nucleotide substitutions, recent research showed that structural variants (SVs), *i.e.* changes in position, presence, and orientation of genomic fragments from a few bases up to several megabases (Mb), represent an important but overlooked aspect of genetic diversity [1]. Long-reads have made it easier to characterize SVs, demonstrating a strong impact on phenotypes, including disease and traits of agronomical interest [2]. Inversions are particularly important for the evolution of biodiversity and are extensively studied in a context of conservation and evolutionary biology [3,4]. Most of those applications, such as genotype-phenotype association and population genomics, requires population studies with many individuals [5], for which the use of long-reads is not cost-efficient while the use of short-reads is not accurate enough, raising the need for alternative type of data. Linked reads combine both the high quality and low cost of short-read sequencing with the long-distance information of long reads. This technology is based on the attachment of barcodes to a long DNA molecule (usually from 5 to 100kb) during library preparation before fragmentation for short-read sequencing [6]. This long-distance information helps scaffolding during genome assembly and identifying large structural changes [7,8]. Several technologies produce linked read data, such as stLFR [6], TELL-seq [9] and, Haplotagging [10]. Happlotagging was optimized for multiplexing and it is thus particularly adequate for large datasets in ecological and population studies.

Structural variants are often studied in two steps: detection and genotyping. Detection aims at characterising SVs present in an individual or population relatively to a reference genome. Genotyping, on the other hand, seeks to determine the presence or absence of each allele of the SVs previously identified in a genome. Detection is generally performed with the highest quality of data, such as long-reads with high depth, which is usually available on a limited number of individuals. Conversely, genotyping needs to be carried out on a larger number of individuals to provide stronger statistical support for analysis and thus uses more cost-effective data such as short-reads [11]. From that point of view, linked-reads thus provides a major advantage by combining a low sequencing cost, allowing to accommodate a large number of individuals, high sequencing quality, and long-range information allowing to capture kb-long molecules spanning over the breakpoints. Several tools exist to detect structural variants with linked-reads data, such as LEVIATHAN [12], Valor [13], Aquila [14], NAIBR [15] or Wrath [16], but there exists no tool for genotyping already discovered structural variants with linked-read data.

This paper presents SVJedi-Tag, the first tool dedicated to inversion genotyping with linked-read data. SVJedi-Tag analyses specific barcode (molecule tags) signals to estimate the presence or absence of an inversion in each individual. We tested our tool on simulated and real haplotagging data from the seaweed fly *Coelopa Frigida*, a species bearing adaptive polymorphic inversions [17].

Materials and methods

Method

Barcode signal for inversion genotyping Linked-reads are produced from long DNA molecules that are individually associated with microbeads coated with unique barcodes. They are then fragmented into barcode-associated short reads [6]. All reads originating from the same DNA molecule share the same barcode. This sequencing technology is therefore characterised by the size of the long DNA molecule (ranging from 1 to 100kb), the number of molecules per barcode (which is close to 1 for haplotagging data) and the average coverage of each molecule by reads (i.e. the number of reads per molecule, usually ranging from 2 to 10). All these characteristics vary depending on the dataset, the quality of the DNA and the library preparation. In linked-read data, the original long molecule is typically covered by a few reads randomly distributed over it [10]. That means that the full molecule cannot simple be re-assembled from its barcoded reads. However, after mapping the reads to a reference genome, long-range information can be recovered from the distribution of similar barcodes (belonging to the same molecule). Hence, to determine if a DNA segment is inverted or not in a sequenced individual, we can extract information from the long molecules that span the inversion breakpoints. SVJedi-tag therefore looks for particular patterns of barcode distributions around the inversion breakpoints to genotype each individual.

SVJedi-Tag is based on three major steps, as shown in Figure 1. First, a variation graph is created from a catalogue of inversions and a reference genome. Then, linked-reads are aligned on this variation graph. Finally, inversions are genotyped by analyzing the barcode distribution around the inversion breakpoints on the graph.

Step 1: Variation graph Instead of mapping the reads to a linear reference genome, we chose to represent the different SV alleles in a variation graph to better represent close and overlapping

variants, following the approach used in the long read genotyper SVJedi-graph [18]. The first step of our method constructs a variation graph (in GFA format) from a catalog of structural variants and a reference genome. In this graph, each node represents a sequence, the edges symbolize the adjacencies between sequences observed in an allele, and each combination of alleles is represented by a path in the graph.

Step 2: Alignment As there is no read-to-graph alignment tool dedicated to linked-read data, we use the short-read mapper VG Giraffe [19] (version 1.43.0), which ignores barcode information during alignment. The alignment process generates a GAF file that contains alignment information for each read on the graph (such as node, position on the node, size, quality, etc.).

Step 3: Genotyping To estimate which allele is present in the sample, we count the amount of barcodes supporting each allele around the inversion breakpoints. For each inversion, we define four regions on the inversion node and its adjacent nodes, in which we will register the observed barcodes: *Left* for the region on the adjacent node at the left of the inversion, *Begin* for the start of the inversion node, *End* for the end of the inversion node, and *Right* for the region on the adjacent node at the right of the inversion. The length of the regions is fixed and is governed by a parameter, whose default value is set to 10 Kb, which is a typical long molecule length (Figure 2).

Based on read alignments, we output a list of barcode found in each of the four regions. A given barcode can belong to three classes of signal: Reference, Alternative and Undetermined. The 'Reference' and 'Alternative' classes are informative because reads which share this same barcode are only found in two regions on either side of an arc of the graph, that is an allele-specific adjacency. The 'Undetermined' class corresponds to barcodes that are found in more than two regions or in both regions of the inversion node and are therefore non-informative. These 'Undetermined' signals result from molecules either including in the inverted segment not spanning any breakpoint, or that are larger than the inversion and span both breakpoints.

Depending on their proximity to other SVs, the adjacent nodes can be smaller than the region length (10 Kb). In this case, a given region can be defined on more than one node, by traversing the graph with a depth-first search so that each different paths has the wanted length (10 Kb). If the inversion node is smaller than twice the region length, then the length of the Begin and End regions is reduced to half the length of the inversion, so that these two regions do not overlap.

Finally, for each inversion, an allelic ratio is calculated as the number of barcodes supporting each allele. The three possible genotypes for a diploid individual are then called using fixed thresholds: below 0.2, the inversion will be genotyped as homozygous reference, above 0.8, homozygous alternative, and between these two thresholds, heterozygous.

Implementation SVJedi-Tag is implemented in Python 3. The code is available on GitHub (https://github.com/Mtemperville/SVJedi-Tag). SVJedi-Tag takes as input data: a file containing inversions in VCF format, which can also include other types of SVs that could impact genotyping (e.g. insertion close to a breakpoint), a linked-reads file (fastq /fastq.gz /fq /fq.gz) and a reference genome (fastq/fa/fna). The output file is a VCF file containing the predicted genotypes for all input inversions.

SVJedi-Tag also produces intermediate files containing the variation graph (GFA) and the alignment file (GAF), which can be used to launch the tool by skipping steps already performed.

Application data

We tested SVJedi-Tag on three types of datasets: (1) simulated linked-reads with simulated inversions; (2) real linked-reads with simulated inversions and (3) real linked-reads with a real 3 Mb known inversion. For the real linked-read data, we used haplotagging data from 14 individuals of *Coelopa frigida* with different sequencing depths ranging from 1X to 7X and the associated reference genome of *Coelopa frigida* (250Mb).

We simulated inversions in the *Coelopa frigida* reference genome, restricted to the first five chromosomes (190Mb). For test (1), we first sampled uniformly 60 non-overlapping 50kb-long segments along the genome. Then we generated 2 haploid *C. frigida* genomes using Visor hack [20], by inverting segments from two different subsets of the 60 segments: 20 segments were inverted in both haplotypes (expected homozygous alternative genotype), 20 other segments were inverted in only one of the two haplotypes (expected heterozygous genotypes). The 20 remaining non-inverted segments represent expected homozygous reference genotypes. We then use both haploid genomes to simulate linked-reads with Visor Xenia [20] using several linked-read parameters inspired from the the characteristics of the real haplotagging data (estimated using the `LR_Stat` script included in the SVJedi-Tag repository).

For test (2), we randomly determined the positions of 40 inversions and modified the *C. frigida* reference genome by inverting half of them in the genome using Visor Hack. Real individuals are thus expected to be homozygous alternative for the 20 inversions put in the reference genome and homozygous reference for the 20 other inversions. We repeated this simulation for five fixed inversion lengths: 10 kb, 25 kb, 50 kb, 100 kb, 500 kb and 1 Mb. Finally, for set of inversions, we ran SVJedi-Tag with the VCF containing the 40 inversions of a given length and the empirical linked-reads from 14 *C. frigida* samples.

For test (3), to assess the accuracy of the method on real data and real inversion, we genotyped a known 3 Mb inversion of *Coelopa Frigida* [17] located on chromosome LG4 (*Cf-Inv(4.1)*), with coordinates 1393751-4456396 determined using long-reads, and previously genotyped by PCR in the 14 individuals.

Results

Tests with simulated data and 50kb-long simulated inversions demonstrated that SVJedi-Tag has a high genotyping rate (96%) and a high genotyping accuracy (92%) despite shallow sequencing depth. In fact, we simulated linked-read data with characteristics comparable to real data obtained from *Coelopa frigida*, *i.e.* an average sequencing depth of 2X, an original molecule length of 10kb and a molecule coverage of 0.05 (around 6 reads on average per molecule). We were able to assign a genotype to 57 out of the 60 simulated 50 Kb inversions. The remaining three inversions could not be genotyped due to an unsufficient number of informative barcodes (here, minimal value of 1). Out of the 57 genotyped inversions, 53 were assigned the true genotype, resulting in a genotyping accuracy of 93 % (see also the distributions of the estimated allelic ratio for the three expected genotypes in Figure 3).

Genotyping performances depend on the linked-read sequencing characteristics: the more reads per molecule, the better the accuracy, with 72 % and 100 % accuracy for molecule coverages of 0.02 and 0.3 respectively (Table 1). Genotyping performances also vary with the length of the region in where barcodes are counted, a parameter which is optimal at values close to the molecule length (Table 1).

To evaluate SVJedi-Tag on more realistic datasets, we simulated homozygous inversions in the reference genome (see Methods) and tested genotyping in empirical haplotagging linked-reads from 14 *Coelopa Frigida* samples. For simulated inversions of 50 Kb, at least 38 inversions out of 40 were genotyped in all samples (min genotyping rate of 95 %), and 12 out of 14 samples obtain a 100 % genotyping accuracy (see Figure 4). The other two samples show a lower genotyping accuracy of 95-97% probably explained by their lower sequencing depth ($< 2X$). It is worth noting that the ungenotyped inversions were randomly simulated in low-mappability regions.

When varying the length of inversions to genotype, we obtained 100 % accuracy and rate for inversions larger than 100 Kb for all real haplotagging samples (Figure 4). However, the accuracy drops slightly to 87% for 25kb-long inversions, and strongly for 10kb-long inversions. This pattern is explained by a higher number of ”uninformative” molecules spanning the whole inverted segment when the inversions are shorter, as well as to the noise associated to the low coverage of reads per molecule.

Last, genotyping with empirical haplotagging data a known 3 Mb inversion of *Coelopa frigida* located on LG4 (*CF-Inv-4.1*) showed 100% accuracy for the 14 individuals previously genotyped by PCR.

Discussion

Our results on simulated and empirical data show that SVJedi-Tag performs very well for genotyping large inversions ($>$50kb) using haplotagging linked-reads, including for data with shallow depth (2-5X). This is a real strength to perform future studies on chromosomal inversions in large population datasets as used in genotype-phenotype studies (GWAS) and ecological genomics. In particular, SVJedi-tag is able to genotype Mb-scale inversions with complex breakpoints, a type of variant which is increasingly studied but hardly accessible with standard methods. Indeed, the inversions most studied for their evolutionary or phenotypic impact are often at least several hundred Kb long and contain several genes.

Genotyping accuracy is improved by different parameters, such as a higher sequencing depth, longer DNA molecules, and well-adjusted length of the genotyping regions. Yet, the number of reads per molecule seems to be the parameter with the greatest impact on genotyping accuracy, and further molecular development to enhance this coverage would likely be beneficial to SV detection and genotyping.

Most genotyping errors were due to allelic ratio being just below or above the threshold. A future improvement will therefore consist in relying on a probabilistic model to infer genotype likelihoods instead. Future developments of SVJedi-tag include testing different parameters to improve genotyping of smaller inversions and exploring the impact of the distance between inversions. SVJedi-tag will also be tested on other empirical linked-read datasets with known inversions to better optimize

inversion genotyping in non-model species. Finally, in the future, we would like to extend the genotyping method to other types of structural variants.

Acknowledgements

We thank the GenOuest bioinformatics core facility for providing the computing infrastructure. We are grateful to GenSeq platform (Université de Montpellier) and MGX platform (CNRS Montpellier) for the haplotagging data on *C. frigida*. C. Mérot is supported by an ERC-starting grant *Evol-SV*.

Figures and Tables

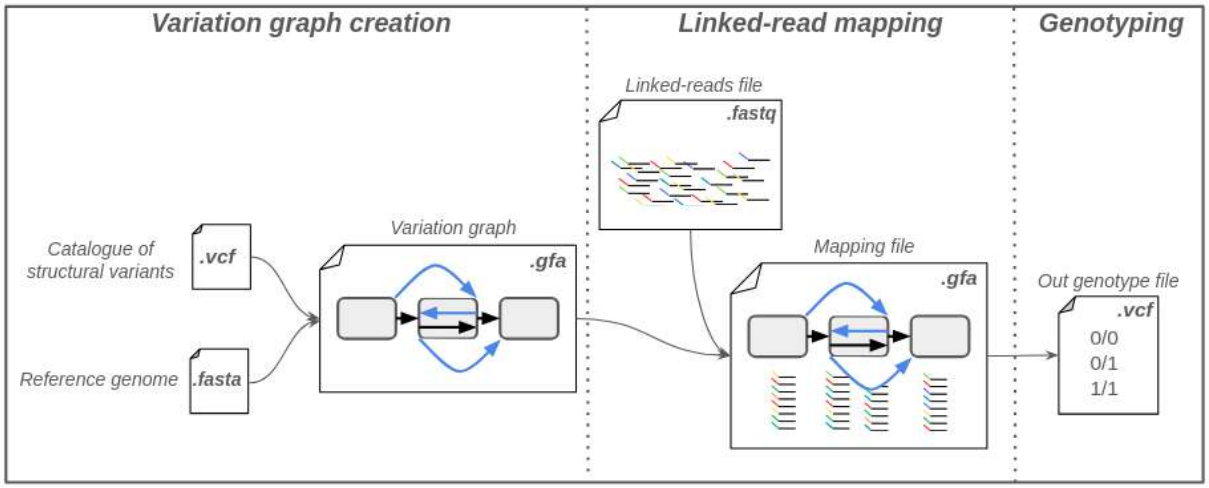


Fig. 1. *Overview of the three main steps of SVjedi-Tag.*

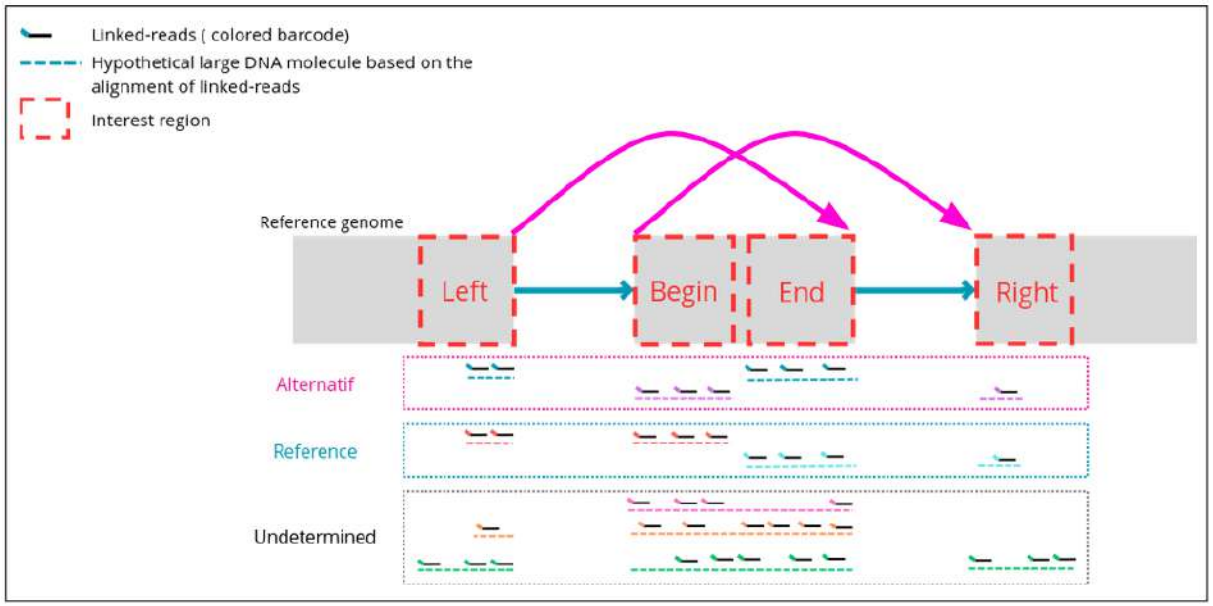


Fig. 2. *Illustration of barcode distributions around the inversion breakpoints. Barcodes with the same colour belong to the same initial DNA molecule. Three different classes of signal can be identified: barcodes supporting the alternative allele with a long molecule spanning the alternative allele adjacencies (in the pink box), barcodes supporting the reference allele with a long molecule spanning the reference allele adjacencies (in the blue box), and non-informative barcodes for which the long molecules span no breakpoints or both breakpoints (in the grey box).*

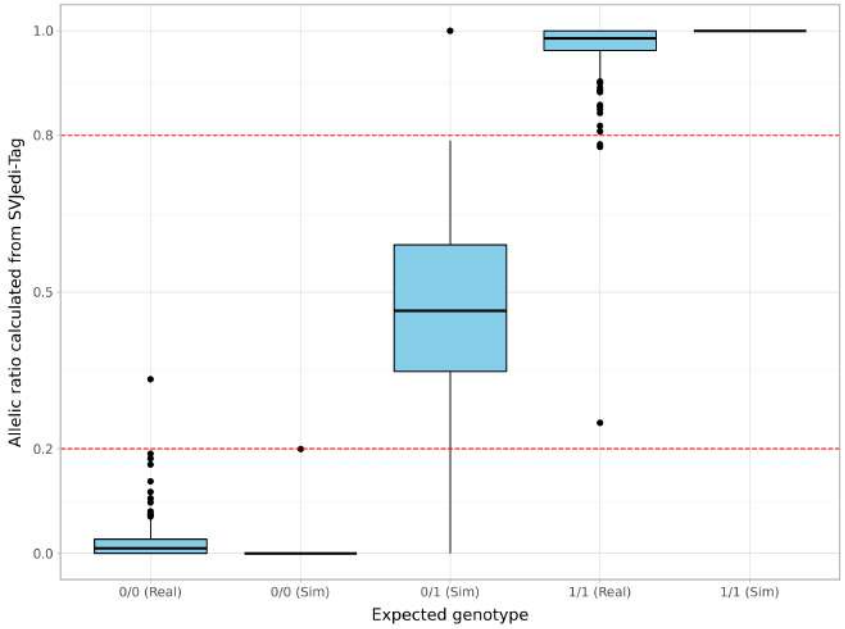


Fig. 3. *Boxplot of estimated allelic ratios predicted by SVJedi-Tag as a function of the expected genotype for inversions of 50 Kb in real and simulated linked-reads. (0/0 : Homozygote reference, 1/1 : Homozygote alternative, 0/1 : Heterozygote) (Sim : simulate linked-reads, Real : real haplotagging linked-reads).*

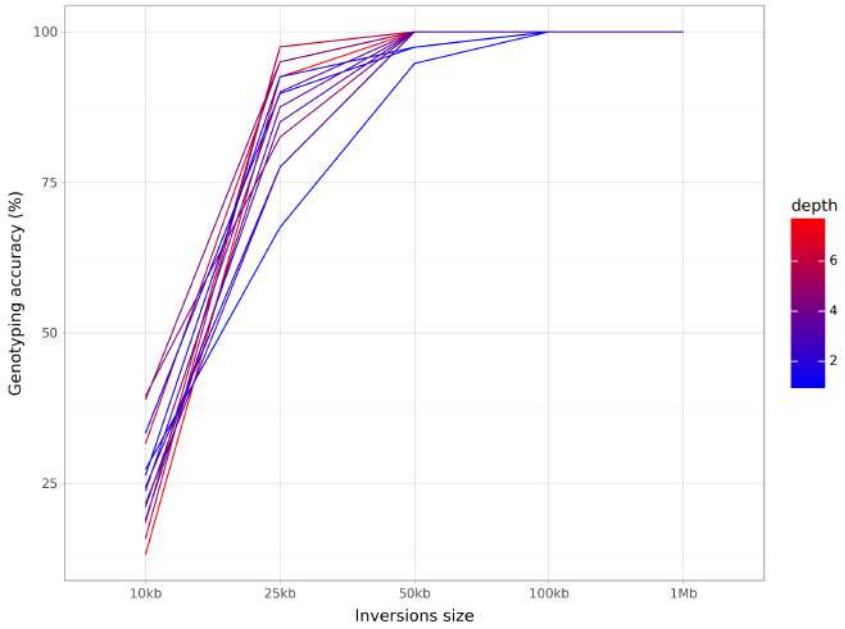


Fig. 4. *Genotyping accuracy for inversions simulated in the reference genome and genotyped in empirical linked-reads from 14 Coelopa frigida samples as a function of inversion length. For each inversion length, 40 inversions were simulated. Each sample is represented by a line whose color reflects sequencing depth, varying between 0.9X and 7.7X.*

	Percentage of molecule covered by reads	Region length parameter value	Genotyping accuracy	Genotyping rate
Molecule size 8 kb	0.02	10 Kb	80%	80%
	0.05	10 Kb	85%	95%
	0.1	10 Kb	89%	96%
	0.2	10 Kb	89%	98%
Molecule size 10 kb	0.02	10 Kb	68%	88%
	0.05	10 Kb	92%	96%
	0.1	10 Kb	94%	98%
	0.2	10 Kb	96%	98%
Molecule size 10 kb	0.05	1 Kb	72%	46%
	0.05	5 Kb	92%	91%
	0.05	15 Kb	93%	98%

Tab. 1. *Accuracy and genotyping rate obtained with SVJedi-Tag on simulated linked-read data, for different simulated molecule length (8 or 10 kb), molecule coverage (0.02, 0.05, 0.1 and 0.2, for each molecule length) and different values for the region length parameter (on data with molecule length of 10 kb and molecule coverage of 0.05).*

References

[1] Mérot C, Oomen RA, Tigano A, Wellenreuther M. A Roadmap for Understanding the Evolutionary Significance of Structural Genomic Variation. Trends in Ecology & Evolution. 2020 Jul;35(7):561-72. Publisher: Elsevier. Available from: https://www.cell.com/trends/ecology-evolution/abstract/S0169-5347(20)30076-8.

[2] Mahmoud M, Gobet N, Cruz-Dávalos DI, Mounier N, Dessimoz C, Sedlazeck FJ. Structural variant calling: the long and the short of it. Genome Biology. 2019 Nov;20(1):246. Available from: https://doi.org/10.1186/s13059-019-1828-7.

[3] Davey JW, Barker SL, Rastas PM, Pinharanda A, Martin SH, Durbin R, et al. No evidence for maintenance of a sympatric Heliconius species barrier by chromosomal inversions. Evolution Letters. 2017 June;1(3):138-54. Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6122123/.

[4] Wellenreuther M, Bernatchez L. Eco-Evolutionary Genomics of Chromosomal Inversions. Trends in Ecology & Evolution. 2018 June;33(6):427-40. Available from: https://www.sciencedirect.com/science/article/pii/S0169534718300788.

[5] Hooper DM, McDiarmid CS, Powers MJ, Justyn NM, Kučka M, Hart NS, et al. Spread of Yellow-Bill-Color Alleles Favored by Selection in the Long-Tailed Finch Hybrid System. Current Biology. 2024 December;34(23):5444-56.e8. Available from: https://linkinghub.elsevier.com/retrieve/pii/S0960982224013733.

[6] Wang O, Chin R, Cheng X, Wu MKY, Mao Q, Tang J, et al. Efficient and unique cobarcoding of second-generation sequencing reads from long DNA molecules enabling cost-effective and accurate sequencing, haplotyping, and de novo assembly. Genome Research. 2019 May;29(5):798-808. Available from: https://genome.cshlp.org/content/29/5/798.

[7] Weisenfeld NI, Kumar V, Shah P, Church DM, Jaffe DB. Direct Determination of Diploid Genome Sequences. Genome Research. 2017 May;27(5):757-67. Available from: http://genome.cshlp.org/content/27/5/757.

[8] Yeo S, Coombe L, Warren RL, Chu J, Birol I. ARCS: Scaffolding Genome Drafts with Linked Reads. Bioinformatics. 2018 March;34(5):725-31. Available from: https://doi.org/10.1093/bioinformatics/btx675.

[9] Chen Z, Pham L, Wu TC, Mo G, Xia Y, Chang PL, et al. Ultralow-input single-tube linked-read library method enables short-read second-generation sequencing systems to routinely generate

highly accurate and economical long-range sequencing information. Genome Research. 2020 June;30(6):898-909. Available from: https://genome.cshlp.org/content/30/6/898.

[10] Meier JI, Salazar PA, Kučka M, Davies RW, Dréau A, Aldás I, et al. Haplotype tagging reveals parallel formation of hybrid races in two butterfly species. Proceedings of the National Academy of Sciences of the United States of America. 2021 June;118(25):e2015005118. Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8237668/.

[11] Duan X, Pan M, Fan S. Comprehensive Evaluation of Structural Variant Genotyping Methods Based on Long-Read Sequencing Data. BMC Genomics. 2022 April;23(1):324. Available from: https://doi.org/10.1186/s12864-022-08548-y.

[12] Morisse P, Legeai F, Lemaitre C. LEVIATHAN: efficient discovery of large structural variants by leveraging long-range information from Linked-Reads data. bioRxiv; 2021. Preprint available under CC BY-ND 4.0 License. Available from: https://www.biorxiv.org/content/10.1101/2021.03.25.437002v1.

[13] Karaoğlanoğlu F, Ricketts C, Ebren E, Rasekh ME, Hajirasouliha I, Alkan C. VALOR2: characterization of large-scale structural variants using linked-reads. Genome Biology. 2020 March;21(1):72. Available from: https://doi.org/10.1186/s13059-020-01975-8.

[14] Zhou X, Zhang L, Weng Z, Dill DL, Sidow A. Aquila enables reference-assisted diploid personal genome assembly and comprehensive variant detection based on linked reads. Nature Communications. 2021 February;12:1077. Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7889865/.

[15] Elyanow R, Wu HT, Raphael BJ. Identifying structural variants using linked-read sequencing data. Bioinformatics. 2018 January;34(2):353-60. Available from: https://doi.org/10.1093/bioinformatics/btx712.

[16] Orteu A, Kucka M, Gordon IJ, Ng'iru I, van der Heijden ESM, Talavera G, et al. Transposable Element Insertions Are Associated with Batesian Mimicry in the Pantropical Butterfly Hypolimnas misippus. Molecular Biology and Evolution. 2024 March;41(3):msae041. Available from: https://doi.org/10.1093/molbev/msae041.

[17] Mérot C, Berdan EL, Cayuela H, Djambazian H, Ferchaud AL, Laporte M, et al. Locally Adaptive Inversions Modulate Genetic Variation at Different Geographic Scales in a Seaweed Fly. Molecular Biology and Evolution. 2021 September;38(9):3953-71. Available from: https://doi.org/10.1093/molbev/msab143.

[18] Romain S, Lemaitre C. SVJedi-graph: improving the genotyping of close and overlapping structural variants with long reads using a variation graph. Bioinformatics. 2023 June;39(Suppl 1):i270-8. Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC10311344/.

[19] Pangenomics enables genotyping of known structural variants in 5202 diverse genomes | Science;. Available from: https://www.science.org/doi/10.1126/science.abg8871.

[20] Bolognini D, Sanders A, Korbel JO, Magi A, Benes V, Rausch T. VISOR: a versatile haplotype-aware structural variant simulator for short- and long-read sequencing. Bioinformatics. 2020 February;36(4):1267-9. Available from: https://doi.org/10.1093/bioinformatics/btz719.

MetagenBERT: a Transformer Architecture using Foundational DNA Read Embedding Models to enhance Disease Classification

Gaspar ROY[1], Eugeni BELDA[1,2], Edi PRIFTI[1,2], Yann CHEVALEYRE[3] and Jean-Daniel ZUCKER[1,2]

[1] IRD, Sorbonne University, UMMISCO, 32 avenue Henry Varagnat, Bondy Cedex, France
[2] Sorbonne University, INSERM, Nutriomics, 91 bvd de l'hopital 75013 Paris, France
[3] LAMSADE, Dauphine University, PSL Research University, Place du Maréchal de Lattre de Tassigny, Paris, France

Corresponding author: gaspar.roy@ird.fr

Keywords Metagenomics, Gut microbiome, Transformer models, DNA sequence embedding, Disease classification

Abstract *Microbial ecosystems constitute complex yet information-rich environments whose characterization is crucial for understanding host health and disease. Among them, the human gut microbiome has emerged as a key "super-integrator", owing to its dense interactions with host physiology and its established associations with a wide spectrum of pathologies. Driven by advances in high-throughput sequencing technologies and the continuous decline in associated costs, metagenomic studies have expanded exponentially, generating massive amounts of sequencing data and opening new avenues for data-driven disease modeling. Conventional approaches to microbiome analysis predominantly rely on the alignment of DNA sequencing reads against reference databases to infer microbial composition and profiling at the species level. While effective, these methods are inherently constrained by reference bias and limited taxonomic resolution. Recent advances in artificial intelligence—particularly in Natural Language Processing (NLP) offer new methodological perspectives for metagenomic data representation. In this study, we present MetagenBERT, a Transformer-based framework to embed metagenomes that relies on the foundational models DNABERT-2 and DNABERT-S for the embedding of DNA sequencing reads. Our approach encodes gut microbiome metagenome in a taxonomy-agnostic manner, enabling direct downstream application to disease classification tasks. We demonstrate that MetagenBERT reaches similar performance to state-of-the-art abundance-based models for cirrhosis prediction and surpasses them in the more challenging context of type 2 diabetes. Furthermore, we introduce an alternative representation of metagenomes based on read-level embeddings aggregated into abundance vectors, demonstrating their complementarity with conventional species-level abundance metrics.*

Introduction

The human gut microbiome, consisting of bacteria, fungi, viruses, archaea, and eukaryotes, outnumbers human cells tenfold [1]. Its diversity and composition are critical indicators of patient health status [2] [3] [4]. Next-Generation Sequencing (NGS) has enabled cost-effective analyses of microbial ecosystems, advancing fields like metagenomics, which profiles the whole DNA from a sample by generating typically millions of short sequences per sample (100 to 300 bases) [5]. While long-read sequencing is emerging [6], allowing to retrieve reads of thousands if not millions of bases, short-read methods remain dominant due to cost-effectiveness and their ability to estimate quantification profiles.

Traditional bioinformatics methods for microbiome sample analysis involves identifying the species it contains, computing their relative abundance tables, and eventually linking the species to health conditions. They mostly rely on large reference catalogs, making them computationally expensive [7]. A typical pipeline is described in Fig. 1. To reduce dependence on reference databases, Deep Learning (DL) techniques—including sequence classification via TetraNucleotide Frequency (TNF) [8][9] or Natural Language Processing (NLP)[10][11][12] can be used, with the objective to automatically identify abstract representations.

Recently, Large Language Models (LLMs) have been increasingly applied in genomics. Foundation models such as Nucleotide Transformer [13] and DNABERT-2 [14] have demonstrated promising performance in various classical gene analysis tasks. More specifically, DNABERT-2 has been further fine-tuned into DNABERT-S [15] using contrastive learning to generate embeddings—mathematical vector representations of sequences—by pulling together sequences from closely related species while pushing apart those from more distant species. This approach enhances sequence-to-species grouping (also known as binning).

Meanwhile, MetaTransformer [16], an attention-based model without pre-training, has exhibited interesting performance in sequence classification tasks, highlighting the potential of transformer-based models in genomics.

In order to link microbiome composition to patient health status, different DL approaches such as PopPhy-CNN [17], MML4Microbiome [18], EnsDeepDP [19] and DeepMicro [20] have been developed to extract abundance features, but still struggle with the high-dimensional, sparse nature of abundance tables, risking overfitting. Solutions like data augmentation [21] or simulation can improve training but may reduce diversity [22]. Some methods bypass abundance tables altogether, using direct sequence embeddings and averaging those by species, as in Metagenome2Vec [23].

However, studies have identified several limitations in using species composition for predicting metagenomic disease associations. First, species detection relies heavily on reference catalogs, which, despite their continuous expansion, still fails to capture a substantial portion of microbiome diversity [24]. Additionally, the choice of taxonomic resolution significantly impacts prediction accuracy. Research has demonstrated that individuals from the same species can exhibit varying, and sometimes even opposing, effects on disease development. Consequently, alternative approaches, such as functional guild-based representations, have been proposed to improve the representation of metagenomic samples [25].

A key challenge in metagenomic analysis is the nature of metagenomic data, which consists of vast amounts of short reads—often numbering in the tens of millions per sample—while datasets typically comprise only a few hundred samples. As a result, metagenomic datasets are highly complex, requiring feature extraction from a limited number of examples. This imbalance makes it difficult to identify meaningful patterns that generalize well for classification tasks while mitigating the risk of overfitting. [26]

For these reasons, we proposed to explore the feasibility an end-to-end and species-agnostic approach, based on Transformers architectures, leveraging powerful embeddings and aggregation techniques for improved microbiome classification and disease prediction. These aggregations also

propose a structure comparable to abundance tables but relying on a species-agnostic approach that captures information different from the classic abundance table.

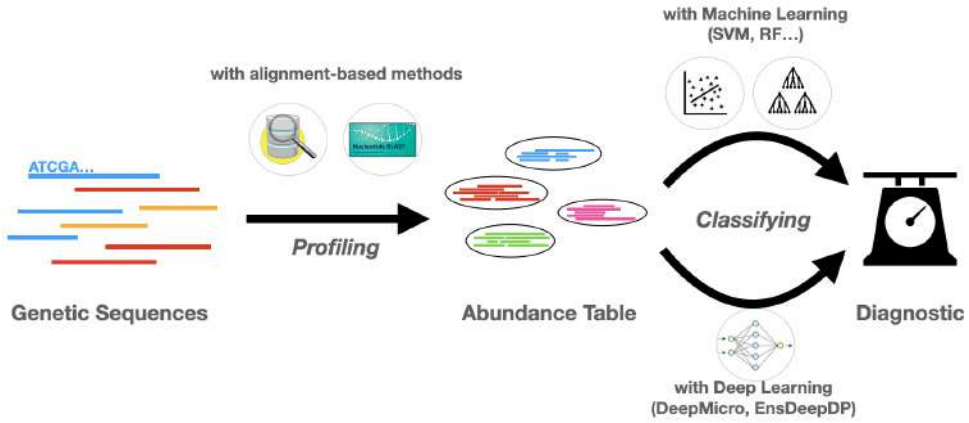


Fig. 1. *Typical bioinformatics pipeline for metagenomics analyses. The widely used approach to perform disease prediction from WGS data is to bin each read into a reference catalog of genes or species. This results in a data table of presence or relative abundance, which can be further used to classify samples in disease groups. This table can then be analyzed to perform classification.*

Materials and Methods

The primary objective of our approach is to develop a method that is both end-to-end and independent of reference catalogs. This method directly processes all DNA reads from a metagenome (after preprocessing) and performs classification without reducing the information to species-level proportions. The underlying idea is that DNA reads contain valuable information beyond species identity, and this information can be effectively extracted using DNA-based large language models.

Datasets

To train and evaluate our method, we used two metagenomic datasets, also employed in the MetaML study [27]. These datasets are related to two clinical conditions, Liver Cirrhosis [28] and Type 2 Diabetes [29]. Both have a comparable proportion of disease and control samples. More information on these datasets can be found in Tab. 1. We downloaded the raw fastq files from EBI (ERP005860) for cirrhosis and NCBI (SRA045646 and SRA050230) for diabetes and cleaned them using fastp with default parameters as illustrated in [30].

	Number of Samples (H/D)	Mean Number of reads per sample	Standard Deviation	Storage	Embedding Time (hours)	Storage of Embeddings	Global Clustering Time
Cirrhosis	232 (114/118)	36,823M	23,396M	1.1T	12h	25.201 T	18h
Type 2 Diabetes	344 (174/170)	34,467M	12;214M	808G	26h	34.977 T	30h

Tab. 1. *Summary of the two datasets used along with computational information*

Pipeline for Metagenome Embedding

Our achitecture consists of several steps detailed in the next subsections. First, the reads from a sample are embedded using a Large Language Model (LLM). Then, depending on the method developed later, this information will either be pooled to form a single vector, the metagenome embedding, or some specific vectors will be selected to represent the metagenome as a smaller bag of vectors, thus drastically reducing its complexity. Keeping in mind that a metagenome is composed of millions of

reads, it is challenging to compress the information from millions of vectors to a single one, both for computational reasons and because of the risk of information loss during aggregation.

Read embedding through Large Language Models The first step of our pipeline consists in transforming the single reads into embeddings. This can be approached by using LLM models trained on DNA data. We used DNABERT-2 [14], a model based on the MosaicBERT [31] architecture, which learns the DNA features by performing Masked Language Modeling on genomic data extracted from various species. For each model, and each sample of our datasets, we embedded the totality of the available sequences, running inference with a batch size of 40000. The embeddings dimension was 768, and the maximum length of a sequence in token was 60 tokens. We then applied average pooling of the embeddings on the token axis, therefore representing the sequences as vectors of dimension 768. Finally, we also used DNABERT-S [15], a fine-tuned version of DNABERT-2 that was explicitly trained to differentiate sequences from different species through contrastive learning, thus learning to generate close embeddings for reads from the same species and more different embeddings for reads originating from different species. Our motivation in focusing in these two architectures, was to obtain both species-specific embeddings with DNABERT-S and more general embeddings with DNABERT-2. All the following steps were applied to both DNABERT-2 and DNABERT-S embedded metagenomes separately. We employed nodes of the Jean-Zay cluster, composed of 96 A100 GPUs to infer the embeddedings of the sequences from both datasets, with a batch size of 40000 reads.

Aggregating Read Information We can consider the read embeddings as local features of our metagenome, we have then used aggregating, sub-sampling and clustering methods in order to represent the global structure of the microbiome.

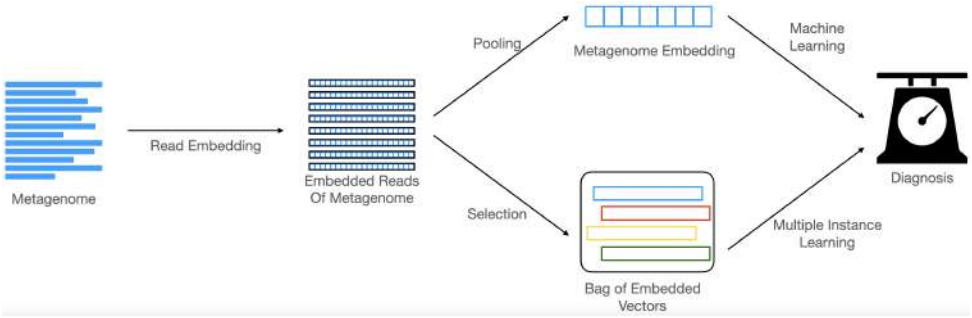


Fig. 2. *The Architecture of our pipeline : metagenomic samples are processed one at a time and their reads embedded by a Transformer model. The metagenome is then represented as a set of embeddings of reads. In order to use it for classification, its dimension is then either reduced to a single vector through aggregation (3), or to a smaller set of vectors through clustering (4) or sampling operations (5). The resulting object can then be used for classification respectively with Machine Learning algorithms or DeepSets [32]*

Simple Aggregation Method As a baseline, we used a simple aggregation method that computes the mean of all embeddings. However, this approach is inherently oversimplifying and leads to significant information loss. For instance, reads with opposing embeddings in certain dimensions may cancel each other out, erasing meaningful variations in the data. Additionally, this method is likely to be biased toward highly represented species, as their read embeddings may cluster more closely together, further limiting the effectiveness of the representation.

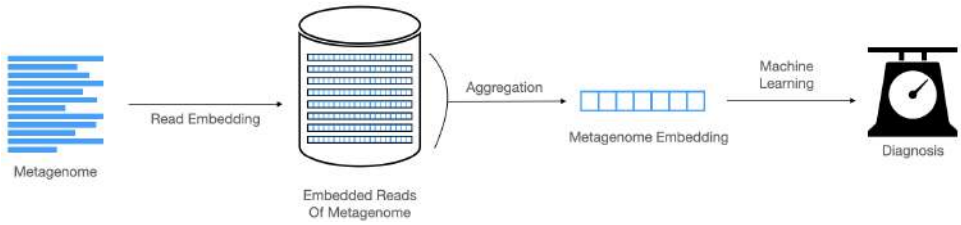


Fig. 3. *The Simple Aggregation Architecture : the set of embedded reads is transformed into one vector by a simple mean operation, then used for classification.*

Clustering Methods In order to reduce the size of our metagenome representation while keeping different types of information, we decided to regroup our read embeddings in clusters, each cluster representing a part of the metagenome. For clustering, we used FAISS Kmeans implementation for rapid, efficient, GPU-optimized clustering [33][34]. We repeated the experiment with a number of clusters ranging from 16 to 16384 (increasing by a factor of 2 each time). The clustering returns two elements : the centroids of each cluster and the assignment of each vector populating the clusters. These assignements can then be used to calculate the proportion of reads in each cluster for each metagenome. We have used this method to develop two different types of clustering pipelines.

Local Clustering When using the method we called "Local Clustering", or MetagenBERT-Local, we performed the KMeans algorithm independantly on each sample of each dataset. A subsample of the reads is used to train the KMeans, then each vector is assigned to its corresponding cluster. The centroids of each cluster are then retrieved to represent the metagenome. In this case, the metagenome is represented by an unordered set of vectors of dimensions number of clusters*embedding size. In this case, each clustering being independent on the sample, calculating the abundance is not useful, for the i-th cluster in the first metagenome is not linked at all to the i-th cluster of the second metagenome. Only the centroids of the clusters are relevant here to represent a metagenome.

Global Clustering The method designated as "Global Clustering", or MetagenBERT-Glob, although also relying on clustering, is different both in its goal and final representation. The idea here is to create a new abundance table representing the microbiome, but based on our embeddings instead of being based on species. To achieve this, we need the different clusters to represent the same parts of the microbiome across all samples from a dataset. To do so, we train our clustering method with reads from 90% of the samples in the dataset (and leave the 10 last percent as holdout). We use 240,000 reads from each sample in the cirrhosis dataset and 180,000 from each sample in the Type 2 Diabetes dataset (for memory considerations), so approximately 1% per sample. This ensures a good enough coverage from every sample in the dataset. Once the KMeans is trained, we take each sample one by one and assign all its reads to their corresponding cluster using nearest neighbors. In this situation, as opposed to local clustering method, samples are not represented by their centroids (as those are common to the whole dataset), but rather by a unique vector : the abundance of each cluster.

Combining Global Clustering Abundance with Species Abundance In order to compare our abundance vector to species-based abundance, we retrieved the abundance vectors of each sample in the dataset and trained a model for disease prediction with the following configurations : by using species-

abundance alone, by using our clustering abundance alone for each number of clusters and by using a concatenation of both for each number of clusters. This experiment allows to compare our method to the baseline, but also to see if the combination of both information makes the results better. If so, this might mean that both methods extract different features from the metagenome.

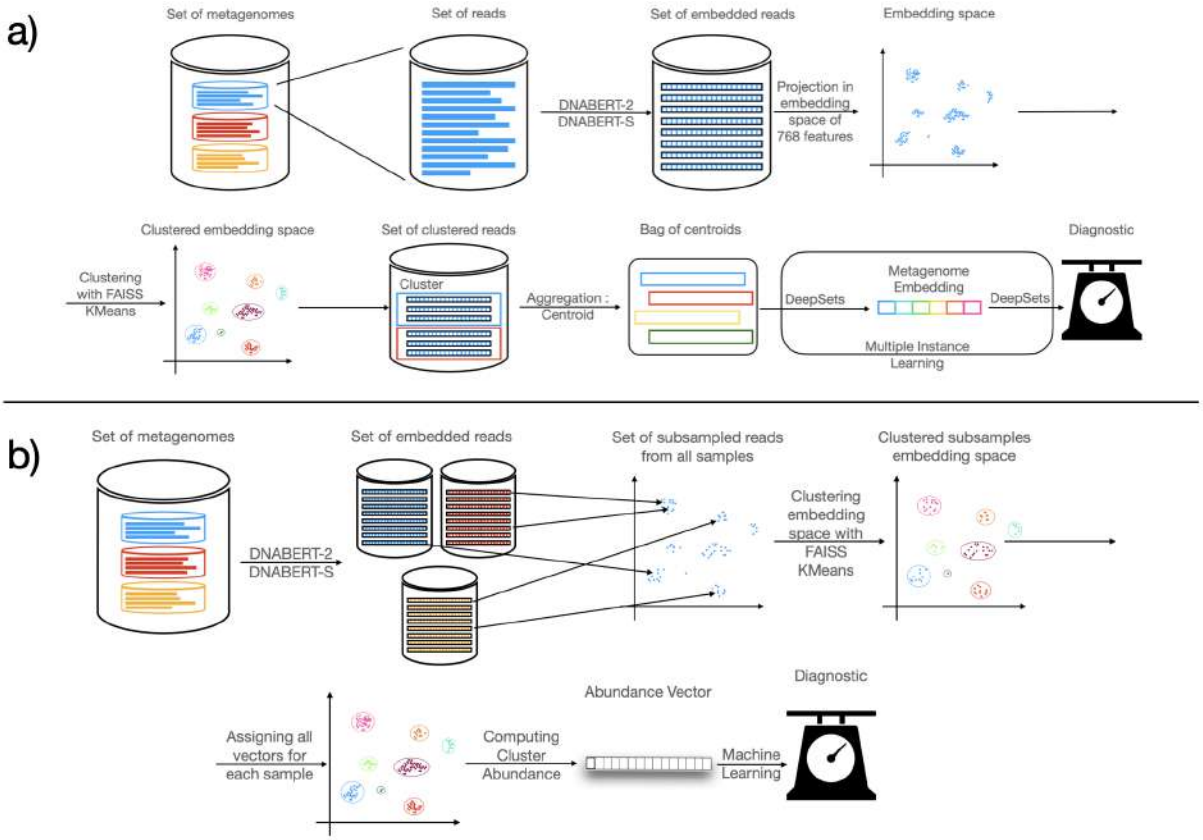


Fig. 4. *The Clustering Architectures : a) represents the Local Clustering method : each sample is independently considered as a set of embedded reads and clusterized using KMeans algorithm from FAISS implementation. The centroid of each cluster is then kept as a representative to produce a new subsample bag that is classified with DeepSets. b) represents the Global Clustering method : a subsample of each embedded metagenome is used to train a KMeans common to all the chosen dataset. Every embedding of each sample is then assigned to its cluster, thus creating a new abundance vector based on embeddings rather than species used for classification.*

Subsampling Method The subsampling method is a simple baseline method that we tested in order to make sure our local clustering method was relevant. To prove that our clusters were significant, and their centroids were good representatives of the different parts of our metagenome, we also tried selecting random reads from each of our samples, choosing as many as we had clusters and treating these reads as if they were the centroids of our clusters. A sample is then represented by a set of vectors and its dimensions are number of samples * embedding size. We expect to obtain worse results with this method than with the local clustering method, thus showing that clustering efficiently captures different relevant parts of our samples.

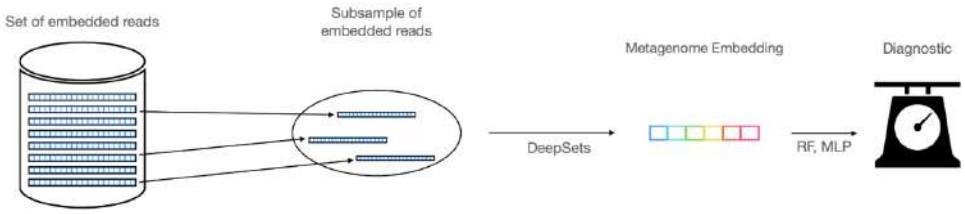


Fig. 5. *The Subsampling Method : a random subsample is drawn from the set of reads, thus obtaining a smaller set that we classify using a DeepSets network.*

Classification from Metagenome Embedding

Once the aggregation methods have been applied, the resulting embedding can be utilized to train a classification model. Two distinct scenarios can be considered based on the representation of the metagenome.

— **Single-Vector Representation:** When employing Simple Aggregation or Global Clustering, the metagenome is represented as a single feature vector. This allows for the application of standard machine learning algorithms, such as Lasso regression, Random Forests, or Multi-Layer Perceptrons (MLP) [35].

— **Set-Based Representation (Multiple Instance Learning):** On the other hand, the remaining aggregation methods, local clustering and subsampling, yield a representation in which the metagenome is characterized as an unordered set of feature vectors. This formulation aligns with the Multiple Instance Learning (MIL) framework [36], wherein a sample is represented by a collection of instances rather than a single feature vector.

To address this MIL problem, we adopt the DeepSets architecture [32]. DeepSets consists of two neural networks, ϕ and ρ, separated by a permutation-invariant pooling layer. The first network, ϕ, processes each vector independently to extract relevant features. These extracted features are then aggregated using the pooling layer, producing a global representation of the metagenome. The second network, ρ, subsequently analyzes this global representation and performs the final disease classification.

In both cases, we performed 10-fold cross validation and computed the standard classification error for each experiment.

Global Clustering Pipeline with 10% data

A metagenomic sample consists of a vast collection of diverse reads, and our approach necessitates a continuous balance between representativity and information aggregation. Embedding and clustering all reads within each metagenomic sample demand substantial computational time and resources. To improve efficiency and reproducibility, we explored a streamlined version of our method by utilizing only a fraction of the data (10%) to assess whether comparable results could be achieved relative to the full dataset : our global clustering model is trained using less than 10% of the available data—specifically, 240,000 reads per sample for the cirrhosis dataset and 180,000 reads per sample for the diabetes dataset. Subsequently, only 10% of the reads from each sample are assigned to the resulting clusters, and cluster proportion vectors are computed based on this subset, rather than on the full set of reads.

Clusters Taxonomic Analysis

Our results suggest that the clusters identified by our method capture distinct dynamics compared to those inferred from species composition. To further investigate this, we analyzed the content of clusters in the T2D dataset derived from DNABERT2 under the global clustering setting with 2048 clusters. For each sample, we generated FASTA files for the 2048 clusters by mapping read assignments back to the original sample FASTA files. We then used Centrifuge v1.0.4 to taxonomically bin individual cluster reads to reference genome sequences from the UHGC1.0 catalog. Reads assigned to species-level representative genomes were retained (min. 10 reads), from which cluster shannon entropy was computed as:

$$\sum(probGenomeBin \times \log_2(probGenomeBin)$$

Where probGenome Bin is the ratio of the number of reads assigned to the Genome by Centrifuge divided by the total number of reads assigned to species-level representative genomes. From this, the mean value of the entropy for the 2048 clusters in each sample were retained as metric describing the complexity of the sample in terms of sequence embeddings.

Results

Baseline Classification using Simple Aggregation

As a baseline, we performed classification on cirrhosis and diabetes datasets using the simple mean of all vector representations. The results, presented in Tab. 2, indicate that while this approach provides some insights for cirrhosis, its performance is significantly lower compared to state-of-the-art methods. This outcome is expected, as the aggregation process leads to substantial information loss and fails to catch the diversity of the metagenome. Although the performances in classifying Type 2 Diabetes are weaker, they are better than could be expected when considering the difficulty of the task and the State of the Art results.

		Cirrhosis	Type 2 Diabetes
Accuracy	DNABERT-2	80.63% (2.56)	69.91% (2.31)
	DNABERT-S	82.3% (2.44)	70.94% (2.26)
AUC	DNABERT-2	0.8797 (0.0210)	0.7711 (0.0219)
	DNABERT-S	0.8827 (0.0206)	0.7860 (0.0232)

Tab. 2. *Performance of MetagenBERT-Aggreg on cirrhosis and type 2 diabetes using DNABERT-2 and DNABERT-S embeddings. Metrics are Accuracy and AUC, in parenthesis is given the standard classification error*

Clustering methods allow valuable insights on microbiome dynamics

Local clustering methods compete with State-of-the-Art

As shown in Fig. 6, the local clustering method yields results that are competitive with state-of-the-art approaches such as MetaML, PopPhyCNN, DeepMicro, MML4Microbiome, and EnsDeepDP on the cirrhosis dataset. While performance is weaker for a small number of clusters, it improves as the number of clusters increases—enhancing the representativity of the sample until it reaches a plateau.

The highest performance is achieved with 2048 clusters, with an AUC of 0.929 and an accuracy of 89.24%).

For the more challenging task of diabetes prediction, the best results are achieved with a smaller number of clusters (128), generally outperforming other methods. In the best case, our approach attains 80.28% accuracy and an AUC of 0.844. These findings suggest two non-exclusive conclusions. First, the clusters generated by our method provide a different representation from species-based approaches, which may be better suited for difficult classification tasks where species-level information is insufficient. Second, the Transformer-based embeddings of reads, found in our centroids, contain relevant features for classification, supporting the hypothesis that Transformers can capture information relevant to the functional role of sequences in microbiome dynamics.

In general, we see the results obtained when using DNABERT-S are slightly weaker than with DNABERT-2, although very close. We can only assume this means that DNABERT-2 creates more general embeddings in which some information different from the ones used to represent the specie of origin are contained.

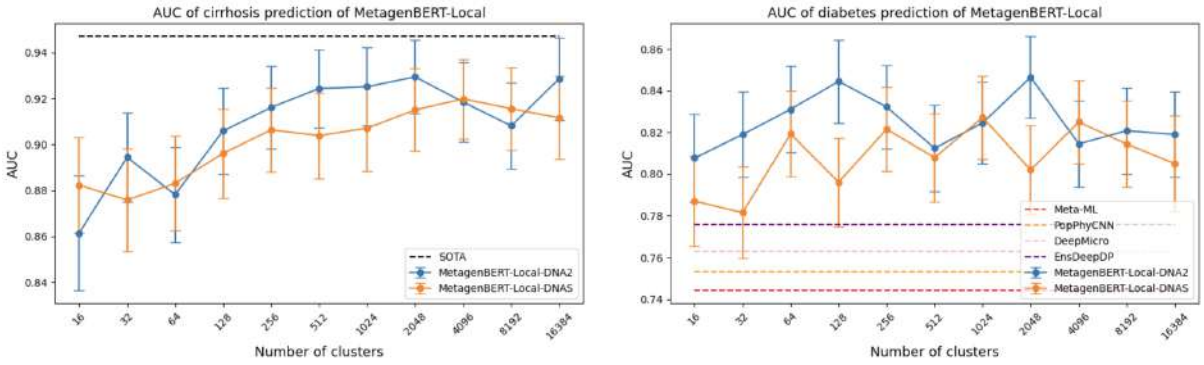


Fig. 6. *Performance Comparison between various SOTA models and our MetagenBERT local clustering classification. Local clustering results in comparable performance to SOTA models on cirrhosis and outperform them on Type 2 Diabetes. For cirrhosis, figure shows the importance of a high number of clusters in prediction quality.*

Global Clustering methods present a different source of information on microbiome dynamics than specie abundance

As previously mentioned, we compared different configurations of our method against state-of-the-art results, including classification using species abundance alone, our global clustering abundance method alone, and a combination of both. We used a simple LASSO classifier, whereas other methods leverage additional sources of information — such as phylogeny in PopPhyCNN [17] and multimodal data in MML4Microbiome [18] — or more complex learning strategies, like CNNs or ensemble learning in EnsDeepDP [19].

Our results indicate that, for a low number of clusters, the global clustering abundance method underperforms due to an insufficient number of features to compete with species-based abundance. However, as the number of clusters increases, it ultimately outperforms species abundance and achieves performance comparable to state-of-the-art methods. These findings suggest that our global clustering approach effectively captures important microbiome dynamics.

Additionally, using abundance features alone, rather than centroids, demonstrates that our method's performance is not solely reliant on the expressive power of Transformer-extracted centroids. Instead, the way reads are distributed within the embedding space also carries valuable information for disease prediction. Lastly, we observe that combining global clustering abundance with species abundance almost always improves classification performance, indicating that these two feature vectors encode different and potentially complementary information.

We therefore conclude that MetagenBERT-Glob may represent a novel source of insight into microbiome data and can be used alongside species abundance to enhance our understanding of the differences between healthy and diseased states.

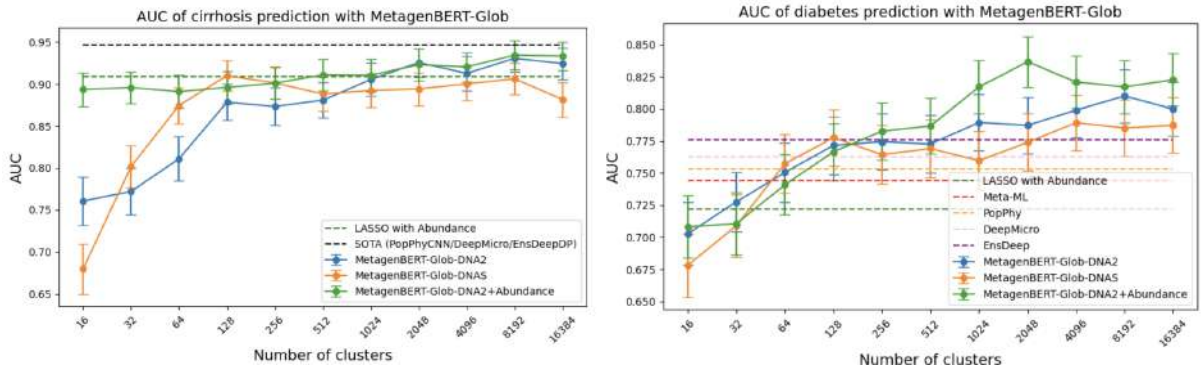


Fig. 7. *Performance Comparison between various SOTA models and Abundance, MetagenBERT-Glob and the concatenation of both with LASSO model. MetagenBERT-Glob results in better performance than Abundance alone and comparable to SOTA models when using a high enough number of clusters. Combining both abundance and global clustering results in better performance than each by themselves, supporting the idea that both carry different types of information. DNABERT-S embeddings result in better results than abundance alone, but fall short to DNABERT-2 performances, especially for the harder task of Diabetes prediction*

Classification with a subsample of reads

We compare the results obtained using subsamples from our cirrhosis dataset to those obtained with the local clustering approach to assess the relevance of the clustering step. The subsampling method generally achieves poor results when using a small number of subsamples but improves as the number of reads increases. This is due to the fact that, when using a low number of samples,, there is a high risk of omitting important regions of the metagenome or overemphasizing non-relevant parts when the number of subsamples, this risk decreases when the number of subsamples increases.

A similar dynamic is observed in the local clustering results. However, the performance achieved with subsampling remains nonetheless lower, acknowledging the importance of the clustering step in our approach.

MetagenBERT-Glob remains accurate even when using a lower amount of data

As shown in Fig. 9, MetagenBERT-Glob achieves robust performance even when applied to a fraction of the available read embeddings, without a substantial decrease in AUC. Across both datasets, and regardless of whether DNABERT-2 or DNABERT-S embeddings are used, results remain stable for every number of clusters. This resilience can be attributed to the compositional nature of the resulting representation: given the high number of reads per metagenomic sample, a representative subset is

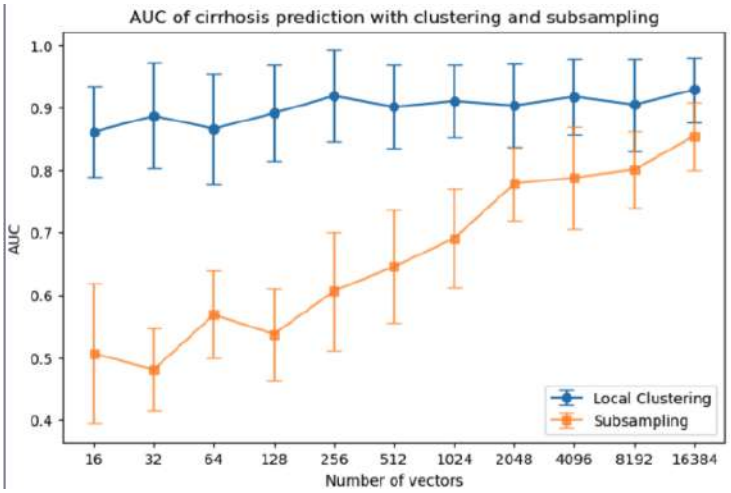


Fig. 8. *Performance Comparison between the local clustering method and the subsampling method. Although the performance increases with the size of the subset, we see that the local clustering method still outperforms this method, underlining the relevance of the clustering process*

sufficient to preserve the integrity of the embedding. This property significantly enhances the scalability of our approach and reduces computational demands, making it more practical for large-scale metagenomic analyses.

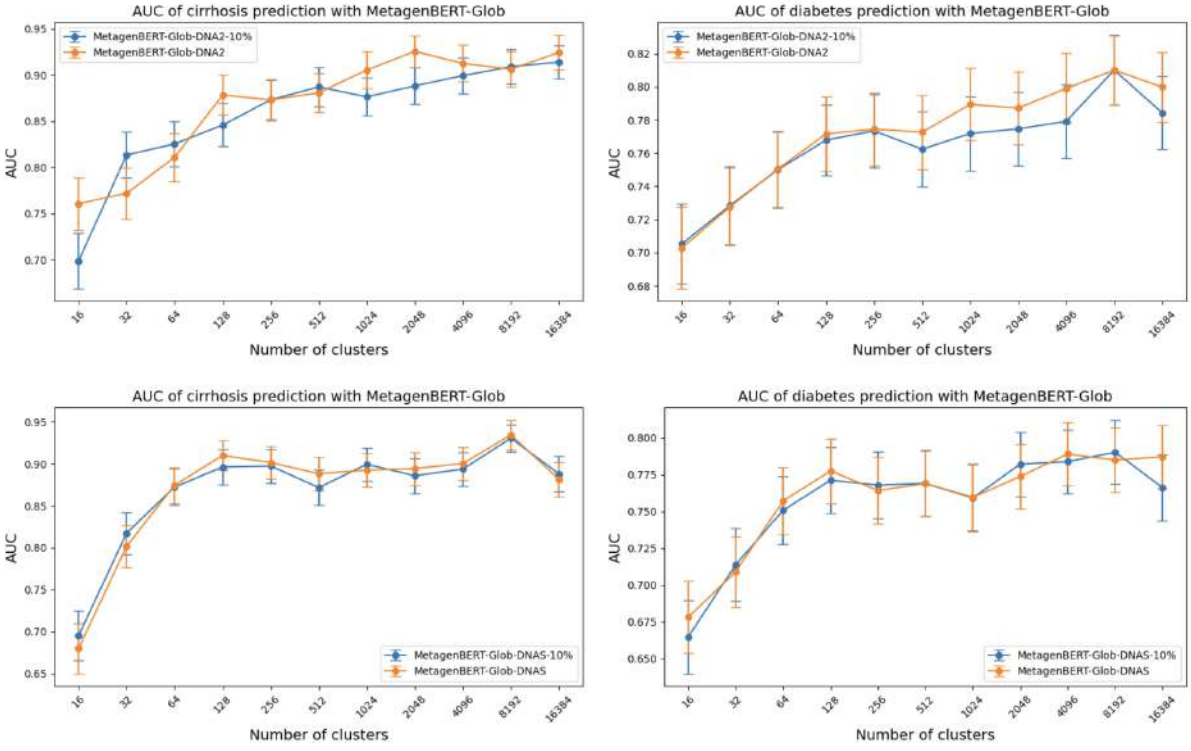


Fig. 9. *Comparison of Prediction performance when using MetagenBERT-Glob with all reads assigned and with only 10% assigned*

Clusters analysis on a taxonomic and functional level

As we can see on Fig. 10, analysis of the clusters identified by our MetagenBERT-Glob algorithm reveals variations in the mean entropy of individual clusters across different groups, exhibiting trends

similar to those observed in the number of genomes identified using Centrifuge. Furthermore, a proportional relationship appears to exist between cluster entropy and the number of genomes retrieved within a sample. These observations support the notion that the clusters generated by our method capture meaningful microbiome dynamics, potentially reflecting underlying diversity and functional characteristics.

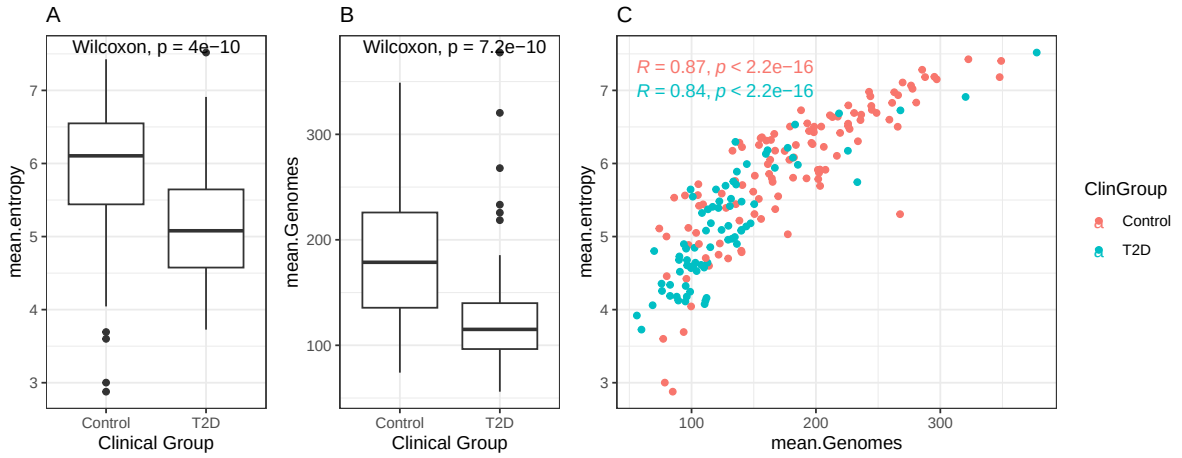


Fig. 10. *Analysis of Clusters composition obtained when using MetagenBERT-Glob with DNABERT-2 embeddings and 2048 clusters on Type 2 Diabetes (T2D) dataset. A) Mean entropy of clusters depending on their group. B) Mean number of genomes retrieved. C) Scatterplot of samples per group comparing their mean entropy and number of genomes retrieved*

Discussion

In this study, we introduced **MetagenBERT**, an end-to-end, species-agnostic framework leveraging foundational DNA language models to classify diseases from raw metagenomic reads. Our results demonstrate that embedding millions of DNA reads with DNABERT-2 and DNABERT-S, followed by tailored aggregation strategies, yields classification performance comparable to or exceeding state-of-the-art models, especially in complex cases such as type 2 diabetes. Notably, the *local clustering* approach captures latent structure in the metagenome that appears more predictive than species abundance alone, suggesting that transformer-derived embeddings encode biologically meaningful features beyond taxonomy—potentially reflecting functional guilds, compositional biases, or strain-level genomic patterns.

Meanwhile, the MetagenBERT-Glob method constructs alternative abundance profiles that complement traditional species-based ones, with performance improving as the number of clusters increases. This synergy between taxonomic and embedding-based representations hints at orthogonal and biologically relevant information being captured by the transformer space. Importantly, the observed gains in prediction are not solely due to the expressive power of DNABERT-derived centroids, but also from the topology of the embedding space itself, as shown by the effectiveness of cluster-based abundance vectors.

While promising, our approach comes with limitations: the computational cost of embedding and clustering tens of millions of reads remains significant - although our results with only a fraction of the data show encouraging possibilities in simplifying the process by not embedding all the reads. More-

over, the interpretability of transformer embeddings is limited, and the datasets used—although standard—comprise relatively few samples, which could affect generalization. Future work will focus on scaling the method, improving biological interpretability (e.g. deeper functional annotation of clusters), and extending the framework to other microbiome types or multi-omics integration.

While Transformer-based models for read embedding, such as DNABERT-2, have demonstrated impressive performance, the field is rapidly evolving, with novel architectures and pretraining strategies emerging regularly that may yield even more informative embeddings. Notably, DNABERT-2 is trained on a diverse set of species, including humans, various animals, viruses, and fungi. This broad training scope, while valuable, may not optimally capture the unique characteristics of metagenomic sequences. Thus, there is potential in developing a model pre-trained specifically on metagenomic datasets, which could provide embeddings better suited for this domain.

We emphasize the complexity of our embedding space: each metagenome consists of tens of millions of reads, embedded here in a 768-dimensional vector space. For clustering, we initially employed the K-Means algorithm due to its simplicity and scalability. However, K-Means relies on Euclidean distance, which may encounter some issues in high-dimensional spaces due to the curse of dimensionality [37]. This phenomenon often causes distance metrics to lose their discriminative power, potentially impairing clustering performance. To address this limitation, we propose exploring more suitable alternatives for high-dimensional clustering. In particular, HDBSCAN [38], a hierarchical density-based algorithm, offers improved sensitivity to varying local densities and is generally more robust than K-Means in complex data landscapes. Additionally, subspace clustering techniques such as CLIQUE or PROCLUS [39] [40] may uncover structure within meaningful low-dimensional subspaces of the embedding space. Spectral clustering [41] also presents a compelling approach, especially when preceded by dimensionality reduction techniques such as PCA or UMAP. However, our tests in reducing dimension with PCA showed low variance when dividing dimension by a factor of 3 to 12, suggesting the dimension reduction loses many valuable features and information.

Furthermore, in the case of MetagenBERT-Local, we decided to address the Multiple Instance Learning problem with the DeepSets architecture. We want to point out that this architecture can face some issues like aggregation bottlenecks or lack of pairwise interaction. To better capture the structure of our clustered space, we suggest using more modern networks such as Set Transformers [42], Graph Neural Networks [43] and PointNet++ [44].

Additionally, we envision that **MetagenBERT** could serve as a building block for more interpretable and functionally grounded microbiome diagnostics, particularly in clinical contexts where species-level methods fall short. Overall, our results suggest that language-model-based embeddings represent a novel and promising axis of representation for metagenomic data, capable of enriching or even transcending conventional microbiome analysis pipelines.

Acknowledgments

This work was supported by a grant from the French ”Agence Nationale de la Recherche” (ANR) for the DeepIntegrOmics project number ANR ANR-21-CE45-0030. This work was granted access to the HPC resources of IDRIS under the allocations 2023-AD011014580, 2024-AD011014580R1 and 2024-AD011015723R1 made by GENCI

References

[1] Ferranti EP, Dunbar SB, Dunlop AL, Corwin EJ. 20 Things You Didn't Know About the Human Gut Microbiome. Journal of Cardiovascular Nursing. 2014 Nov;29(6):479-81. Available from: https://journals.lww.com/00005082-201411000-00004.

[2] Pflughoeft KJ, Versalovic J. Human Microbiome in Health and Disease. Annual Review of Pathology: Mechanisms of Disease. 2012 Feb;7(1):99-122. Available from: https://www.annualreviews.org/doi/10.1146/annurev-pathol-011811-132421.

[3] Le Chatelier E, Nielsen T, Qin J, Prifti E, Hildebrand F, Falony G, et al. Richness of human gut microbiome correlates with metabolic markers. Nature. 2013 Aug;500(7464):541-6. Available from: https://www.nature.com/articles/nature12506.

[4] Qin N, Yang F, Li A, Prifti E, Chen Y, Shao L, et al. Alterations of the human gut microbiome in liver cirrhosis. Nature. 2014 Sep;513(7516):59-64. Available from: http://www.nature.com/articles/nature13568.

[5] MetaHIT Consortium, Qin J, Li R, Raes J, Arumugam M, Burgdorf KS, et al. A human gut microbial gene catalogue established by metagenomic sequencing. Nature. 2010 Mar;464(7285):59-65. Available from: http://www.nature.com/articles/nature08821.

[6] Oehler JB, Wright H, Stark Z, Mallett AJ, Schmitz U. The application of long-read sequencing in clinical settings. Human Genomics. 2023 Aug;17(1):73. Available from: https://humgenomics.biomedcentral.com/articles/10.1186/s40246-023-00522-3.

[7] AltschuP SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic Local Alignment Search Tool:8.

[8] Sharma D, Paterson AD, Xu W. TaxoNN: ensemble of neural networks on stratified microbiome data for disease prediction. Bioinformatics. 2020 Nov;36(17):4544-50. Available from: https://academic.oup.com/bioinformatics/article/36/17/4544/5843784.

[9] Nissen JN, Johansen J, Allesøe RL, Sønderby CK, Armenteros JJA, Grønbech CH, et al. Improved metagenome binning and assembly using deep variational autoencoders. Nature Biotechnology. 2021 May;39(5):555-60. Available from: http://www.nature.com/articles/s41587-020-00777-4.

[10] Mock F, Kretschmer F, Kriese A, Böcker S, Marz M. BERTax: taxonomic classification of DNA sequences with Deep Neural Networks. Bioinformatics; 2021. Available from: http://biorxiv.org/lookup/doi/10.1101/2021.07.09.451778.

[11] Liang Q, Bible PW, Liu Y, Zou B, Wei L. DeepMicrobes: taxonomic classification for metagenomics with deep learning. NAR Genomics and Bioinformatics. 2020 Mar;2(1):lqaa009. Available from: https://academic.oup.com/nargab/article/doi/10.1093/nargab/lqaa009/5740226.

[12] Roy G, Prifti E, Belda E, Zucker JD. Deep learning methods in metagenomics: a review. Microbial Genomics. 2024 Apr;10(4). Available from: https://www.microbiologyresearch.org/content/journal/mgen/10.1099/mgen.0.001231.

[13] Dalla-Torre H, Gonzalez L, Mendoza-Revilla J, Lopez Carranza N, Grzywaczewski AH, Oteri F, et al. Nucleotide Transformer: building and evaluating robust foundation models for human genomics. Nature Methods. 2025 Feb;22(2):287-97. Available from: https://www.nature.com/articles/s41592-024-02523-z.

[14] Zhou Z, Ji Y, Li W, Dutta P, Davuluri R, Liu H. DNABERT-2: Efficient Foundation Model and Benchmark For Multi-Species Genome. arXiv; 2024. ArXiv:2306.15006 [q-bio]. Available from: http://arxiv.org/abs/2306.15006.

[15] Zhou Z, Wu W, Ho H, Wang J, Shi L, Davuluri RV, et al.. DNABERT-S: Pioneering Species Differentiation with Species-Aware DNA Embeddings. arXiv; 2024. ArXiv:2402.08777 [q-bio]. Available from: http://arxiv.org/abs/2402.08777.

[16] Wichmann A, Buschong E, Müller A, Jünger D, Hildebrandt A, Hankeln T, et al. MetaTransformer: deep metagenomic sequencing read classification using self-attention models. NAR Genomics and Bioinformatics. 2023 Jul;5(3):lqad082. Available from: https://academic.oup.com/nargab/article/doi/10.1093/nargab/lqad082/7269178.

[17] Reiman D, Metwally AA, Sun J, Dai Y. PopPhy-CNN: A Phylogenetic Tree Embedded Architecture for Convolutional Neural Networks to Predict Host Phenotype From Metagenomic Data. IEEE Journal of Biomedical and Health Informatics. 2020 Oct;24(10):2993-3001. Available from: https://ieeexplore.ieee.org/document/9091025/.

[18] Lee SJ, Rho M. Multimodal deep learning applied to classify healthy and disease states of human microbiome. Scientific Reports. 2022 Dec;12(1):824. Available from: https://www.nature.com/articles/s41598-022-04773-3.

[19] Shen Y, Zhu J, Deng Z, Lu W, Wang H. Ensdeepdp: An Ensemble Deep Learning Approach for Disease Prediction Through Metagenomics. IEEE/ACM Transactions on Computational Biology and Bioinformatics. 2022:1-14. Available from: https://ieeexplore.ieee.org/document/9866523/.

[20] Oh M, Zhang L. DeepMicro: deep representation learning for disease prediction based on microbiome data. Scientific Reports. 2020 Dec;10(1):6026. Available from: http://www.nature.com/articles/s41598-020-63159-5.

[21] Mulenga M, Abdul Kareem S, Qalid Md Sabri A, Seera M, Govind S, Samudi C, et al. Feature Extension of Gut Microbiome Data for Deep Neural Network-Based Colorectal Cancer Classification. IEEE Access. 2021;9:23565-78. Available from: https://ieeexplore.ieee.org/document/9319639/.

[22] Sczyrba A, Hofmann P, Belmann P, Koslicki D, Janssen S, Dröge J, et al. Critical Assessment of Metagenome Interpretation—a benchmark of metagenomics software. Nature Methods. 2017 Nov;14(11):1063-71. Available from: http://www.nature.com/articles/nmeth.4458.

[23] Queyrel M, Prifti E, Templier A, Zucker JD. Towards end-to-end disease prediction from raw metagenomic data. Genomics; 2020. Available from: http://biorxiv.org/lookup/doi/10.1101/2020.10.29.360297.

[24] Thomas AM, Segata N. Multiple levels of the unknown in microbiome research. BMC Biology. 2019 Dec;17(1):48. Available from: https://bmcbiol.biomedcentral.com/articles/10.1186/s12915-019-0667-z.

[25] Wu G, Zhao N, Zhang C, Lam YY, Zhao L. Guild-based analysis for understanding gut microbiome in human health and diseases. Genome Medicine. 2021 Dec;13(1):22. Available from: https://genomemedicine.biomedcentral.com/articles/10.1186/s13073-021-00840-y.

[26] Calle ML. Statistical Analysis of Metagenomics Data. Genomics & Informatics. 2019 Mar;17(1):e6. Available from: http://genominfo.org/journal/view.php?doi=10.5808/GI.2019.17.1.e6.

[27] Pasolli E, Truong DT, Malik F, Waldron L, Segata N. Machine Learning Meta-analysis of Large Metagenomic Datasets: Tools and Biological Insights. PLOS Computational Biology. 2016 Jul;12(7):e1004977. Available from: https://dx.plos.org/10.1371/journal.pcbi.1004977.

[28] Qin YFLAea N. Alterations of the human gut microbiome in liver cirrhosis. Nature 513, 59–64. 2014. Available from: https://doi.org/10.1038/nature13568.

[29] Qin J CZea Li Y. A metagenome-wide association study of gut microbiota in type 2 diabetes. Nature 490, 55–60. 2012. Available from: https://doi.org/10.1038/nature11450.

[30] Chen S. Ultrafast one-pass FASTQ data preprocessing, quality control, and deduplication using fastp. iMeta. 2023 May;2(2):e107. Available from: https://onlinelibrary.wiley.com/doi/10.1002/imt2.107.

[31] Portes J, Trott A, Havens S, King D, Venigalla A, Nadeem M, et al. MosaicBERT: A Bidirectional Encoder Optimized for Fast Pretraining.

[32] Zaheer M, Kottur S, Ravanbakhsh S, Poczos B, Salakhutdinov R, Smola A. Deep Sets. arXiv; 2018. ArXiv:1703.06114 [cs]. Available from: http://arxiv.org/abs/1703.06114.

[33] Douze M, Guzhva A, Deng C, Johnson J, Szilvasy G, Mazaré PE, et al. The Faiss library. 2024.

[34] Johnson J, Douze M, Jégou H. Billion-scale similarity search with GPUs. IEEE Transactions on Big Data. 2019;7(3):535-47.

[35] Rumelhart GEH David E, Williams RJ. Learning Internal Representations by Error Propagation. MIT Press. 1986.

[36] Babenko B. Multiple Instance Learning: Algorithms and Applications.

[37] G T. A Problem of Dimensionality: A Simple Example. IEEE Transactions on Pattern Analysis and Machine Intelligence. 1979.

[38] Malzer C, Baum M. A Hybrid Approach To Hierarchical Density-based Cluster Selection. In: 2020 IEEE International Conference on Multisensor Fusion and Integration for Intelligent Systems (MFI); 2020. p. 223-8. ArXiv:1911.02282 [cs]. Available from: http://arxiv.org/abs/1911.02282.

[39] Automatic subspace clustering of high dimensional data for data mining applications.

[40] Aggarwal CC, Procopiuc C. Fast algorithms for projected clustering.

[41] Ng AY, Jordan MI, Weiss Y. On Spectral Clustering: Analysis and an algorithm.

[42] Lee J, Lee Y, Kim J, Kosiorek AR, Choi S, Teh YW. Set Transformer: A Framework for Attention-based Permutation-Invariant Neural Networks. arXiv; 2019. ArXiv:1810.00825 [cs]. Available from: http://arxiv.org/abs/1810.00825.

[43] Zhou J, Cui G, Hu S, Zhang Z, Yang C, Liu Z, et al. Graph neural networks: A review of methods and applications. AI Open. 2020;1:57-81. Available from: https://linkinghub.elsevier.com/retrieve/pii/S2666651021000012.

[44] Qi CR, Yi L, Su H, Guibas LJ. PointNet++: Deep Hierarchical Feature Learning on Point Sets in a Metric Space. arXiv; 2017. ArXiv:1706.02413 [cs]. Available from: http://arxiv.org/abs/1706.02413.

Session 2: Metagenomics, Metatranscriptomics, and Microbial Ecosystems Statistics

Metatranscriptomic classification in the study of microbial translocation

Antonin COLAJANNI[1,2], Raluca URICARU[2], Rodolphe THIÉBAUT[1], and Patricia THEBAULT[2]

1 Univ. Bordeaux, INSERM, INRIA, BPH, U1219, F-33000 Bordeaux, France

2 Univ. Bordeaux, CNRS, Bordeaux INP, LaBRI, UMR 5800, F-33400 Talence, France

Corresponding Author: antonin . colajanni @ u-bordeaux.f r

Keywords
Metatranscriptomic, Metagenomic, Microbial translocation, Public data, Benchmark

Abstract

Microbial translocation occurs when bacteria migrate from the gut to the blood due to gut barrier alterations, potentially triggering persistent immune activation and affecting immune responses. Translocation can be studied using whole blood RNA sequencing techniques to analyze the “metatranscriptome”: all the RNA from bacteria, viruses and fungi from the blood. However, a key challenge is reusing cohort data, which primarily consists of human sequences, to characterize the non-human meta-transcriptome. Cohort studies typically focus on human data while minimizing non-human contamination. From a translocation perspective, these non-human sequences become the focus, requiring human sequences to be filtered out, leaving a small fraction (~2%) to be analyzed. In previous work, Nganou-Makamdop *et al.* (1) used a sequence assembly-based pipeline for translocation analysis. We compared this approach with an assembly-free method and found that integrating both strategies into a “hybrid” pipeline improved classification performance in simulations. While real-data validation remains challenging, our results suggest this hybrid strategy enhances microbial translocation analysis.

(1) Nganou-Makamdop K, Talla A, Sharma AA, Darko S, Ransier A, …, Douek DC. Translocated microbiome composition determines immunological outcome in treated HIV infection. Cell. 2021 Jul 22;184(15):3899-3914.e16

OneNet—One network to rule them all: Consensus network inference from microbiome data

Camille CHAMPION[1], Raphaëlle MOMAL[1], Emmanuelle LE CHATELIER[1] , Mathilde SOLA[1], Mahendra MARIADASSOU[2] and Magali BERLAND[1]

1 Université Paris Saclay, INRAE, MGP, 78350, Jouy-en-Josas, France

2 Université Paris Saclay, INRAE, MaIAGE, 78350, Jouy-en-Josas, France

Corresponding Author: magali.berland@inrae.fr

Paper Reference: Champion, Camille, et al. "OneNet—One network to rule them all: Consensus network inference from microbiome data." *PLOS Computational Biology* 20.12 (2024): e1012627. https://doi.org/10.1371/journal.pcbi.1012627

Keywords

Network inference, Stability selection, Microbial ecology, Microbial guild, Gaussian Graphical Models

Abstract

Modeling microbial interactions as sparse and reproducible networks is a major challenge in microbial ecology. Direct interactions between the microbial species of a biome can help to understand the mechanisms through which microbial communities influence the system. Most state-of-the art methods reconstruct networks from abundance data using Gaussian Graphical Models, for which several statistically grounded and computationally efficient inference approaches are available. However, the multiplicity of existing methods, when applied to the same dataset, generates very different networks. In this article, we present OneNet, a consensus network inference method that combines seven methods based on stability selection. This resampling procedure is used to tune a regularization parameter by computing how often edges are selected in the networks. We modified the stability selection framework to use edge selection frequencies directly and combine them in the inferred network to ensure that only reproducible edges are included in the consensus. We demonstrated on synthetic data that our method generally led to slightly sparser networks while achieving much higher precision than any single method. We further applied the method to gut microbiome data from liver-cirrhotic

patients and demonstrated that the resulting network exhibited a microbial guild that was meaningful in terms of human health.

Highlight

Exploring how microbes interact within our intestines is a fascinating challenge at the interface of microbial ecology, bioinformatics, and biostatistics. These interactions shape microbiota composition and influence host health, yet inferring microbial association networks remains methodologically challenging. Each year, new computational approaches for network inference are proposed, yet different methods applied to the same dataset often yield inconsistent results, with no consensus on the best approach.

Our study introduces *OneNet* a new approach that combines seven methods to create a unified, more robust microbial network. On simulated data, *OneNet* demonstrated improved accuracy and reduced network complexity compared to individual methods. Applied to real gut microbiome data from liver cirrhosis patients, our approach identified a cirrhotic cluster—composed of bacteria associated with degraded clinical status—highlighting its potential to better understanding of the role of the gut microbiota on health.

By bridging statistical modeling, validated on both simulated and experimental data, and biological interpretation, this work contributes to advancing microbiome network-based research. It offers a flexible framework that can be leveraged by the JOBIM community to enhance our understanding of host-microbiota interactions and allows applications to other high-dimensional omics datasets.

Fast answers to simple bioinformatics needs and capacity building in an island context, a focus on microbial omics data analysis

Isaure QUETEL [1], Sourakhata TIRERA [2], Damien CAZENAVE [1], Nina ALLOUCH [1], Chloé BAUM [3], Yann REYNAUD [1], Degrâce BATANTOU MABANDZA [1], Virginie NERRIERE [1], Serge VEDY [1], Matthieu POT [1], Sébastien BREUREC [1,4,5,6,7], Anne LAVERGNE [2], Séverine FERDINAND [1], Vincent GUERLAIS [1] and David COUVIN [1,8]*

1 Transmission, Reservoir and Diversity of Pathogens Unit, Pasteur Institute of Guadeloupe, 97139 Les Abymes, Guadeloupe, France

2 Laboratoire des Interactions Virus-Hôtes, Institut Pasteur de la Guyane, 97300 Cayenne, Guyane Française, France

3 Biomics technological Platform, Institut Pasteur, 75015 Paris, France

4 Faculty of Medicine Hyacinthe Bastaraud, University of the Antilles, 97110 Pointe-à-Pitre, France

5 INSERM, Centre for Clinical Investigation 1424, 97110 Pointe-à-Pitre, France

6 Department of Pathogenesis and Control of Chronic and Emerging Infections, University of Montpellier, INSERM, 34394 Montpellier, France

7 Laboratory of Clinical Microbiology, University Hospital Centre of Guadeloupe, 971110 Pointe-à-Pitre, France

8 Laboratoire de Mathématiques Informatique et Applications (LAMIA), Université des Antilles, 97110 Pointe-à-Pitre, Guadeloupe, France

*Corresponding Author: dcouvin@pasteur-guadeloupe.fr

Optional: NA

Keywords

Bioinformatics, tutorial, (meta)genomics, long-read sequencing, workflows.

Abstract

Bioinformatics is increasingly used in various scientific works. Large amounts of heterogeneous data are being generated by scientific teams or laboratories for research purposes. Sequencing and other biological data are difficult to interpret and analyze effectively without dedicated and adapted tools. Several software tools have been developed to facilitate handling and analyses of these types of data. The Galaxy project web platform is one of these software tools that allow free access to users and facilitates the use of thousands of bioinformatics tools. Other software tools like Bioconda or Jupyter Notebook make it easier to install tools and their dependencies for bioinformatics scripts or to offer a user-friendly web interface. In addition to these tools, we can mention RStudio which is an integrated development environment (IDE) facilitating the use

of R scripts. The aim of this study is to provide some guides (or helpers) to the scientific community to perform some bioinformatics or biostatistics analyses in a simpler manner. We also try with this work to democratize well-documented software tools to make them suitable for both bioinformaticians and non-bioinformaticians. We believe that user-friendly guides and real-life/concrete examples will provide end users with suitable and easy-to-use methods for their bioinformatics analysis needs. Furthermore, tutorials and examples of use will be available on our dedicated GitHub repository (https://github.com/karubiotools/AnssBin). These tutorials/examples (in English and/or French) could be used as pedagogical tools promoting bioinformatics analyses and potential answers to some bioinformatics needs. Platforms and/or services play an important role in helping scientists with their bioinformatics data analysis work. These facilities are the cornerstone of bioinformatics capacity building in the overseas islands and support the growth of nascent networks such as KaruBioNet.

Sections of the main text

Introduction

Data analysis is a key method requiring constant updates and adaptations. Several bioinformatic studies depend on robust data analyses and statistical methods to draw significant conclusions and allow a clear understanding of the topic studied. Heterogeneous data is collected at a rapid pace in different laboratories around the world. Regarding the broad field of biology, DNA sequencing data represents a significant part of the biological data analyzed. Various areas and approaches can be used to better understand real-life data, such as metagenomics/metabarcoding, genome assembly and annotation, comparative genomics/data visualization, and gene or motif prediction (among others). Bioinformatics can be used to better understand public health issues such as antibiotic resistance in various pathogens (Escherichia coli, Klebsiella pneumoniae, Mycobacterium tuberculosis…) [1,2].

Several software tools have been developed to facilitate handling and analysis of biological data such as SPAdes for genome de novo assembly and Prokka for prokaryotic genome annotation [3,4]. Concerted actions and developments are being performed to improve the usability of various software tools. On the other hand, simple statistical analyses can be performed using the R software. Some tools have been developed and maintained by a wide community of developers. The Galaxy Platform is a good example of an open-source web tool curated by a large community of scientists [5]. Many Galaxy instances are available for free worldwide (e.g. https://usegalaxy.eu/, https://usegalaxy.fr/, https://galaxy.pasteur.fr/, etc…). Furthermore, Galaxy provides free and easy access to thousands of bioinformatics tools. Other software tools like Bioconda (https://bioconda.github.io/) [6] or Jupyter Notebook (https://jupyter.org/) make it easy to use command lines to run bioinformatics scripts

directly through a terminal or using a web interface. Efforts are also being made to better describe software tools and other digital life-science resources. A concrete example is the bio.tools platform (https://bio.tools/) [7].

The UNIX environment is ideal for performing bioinformatics analyses. Virtual machines are important tools allowing simplified use and access to bioinformatics codes. Several computer codes and programs are already available on these operating systems. Although Microsoft Windows is a well-used operating system, it does not allow for a command-line interface as intuitive and flexible as that of Unix/Linux (for which many tools have been developed). The purpose of this work is to provide some guides (tutorials/examples) to people wishing to perform certain bioinformatic or statistical analyses in an easy way. Some examples of bacterial genomics and metagenomics analyses are shown in this study. Our GitHub repository aims to bring together a huge amount of potentially useful information regarding bioinformatics and biostatistics in one place (in French and/or English). With our approach, we also intend to promote capacity-building in bioinformatics and growth of local bioinformatics platforms like KaruBioNet [8].

Methods

Programming languages

Various programming languages were used to develop specific software tools in function of different laboratories' needs. Some languages and tools were better used for statistical analyses (e.g. R language via RStudio). Other languages, such as C/C++ and Java, have been used to build fast-running algorithms. Furthermore, languages such as Perl and Python have been used to develop programs using a wide variety of secondary modules/libraries and to easily manipulate bioinformatics input/output files. A library like Biopython is notably intensively used for biological sequence analysis. Finally, Bash scripting language could be used to easily run programs or codes in the terminal.

Language	Pros	Cons
C/C++	- High performance & efficiency - Fine control over memory - Widely used in system programming & game dev	- Complex syntax - Manual memory management - Harder debugging
Java	- Platform-independent - Strong memory management - Large ecosystem & libraries	- Slower than compiled languages - Verbose syntax - Requires Java Virtual Machine (JVM)
Perl	- Strong text processing capabilities - Versatile and fast scripting - CPAN module library	- Readability issues - Less modern usage - Slower than compiled languages
Python	- Easy to learn & read	- Slower execution speed

	- Extensive libraries - Great for AI, ML, & data science	- High memory usage - Not ideal for mobile development
R	- Best for statistical computing - Rich visualization tools - Strong package ecosystem	- Slower execution speed - Less suitable for general programming - Steeper learning curve for non-statisticians

Tab. 1: *Comparison table of some popular programming languages (C/C++, Java, Python, Perl, and R).*

Workflow management systems

Creating a workflow enables you to define a pipeline, i.e. a set of steps (or processes) that always follow each other in the same order, with the same structure of inputs, outputs and defined parameters. Thus, if someone recovers a workflow from a previous analysis and uses the same identical input data in the workflow, it will produce identical output results (only if its working environment is stable, as mentioned above). An example of a workflow framework is Nextflow (https://www.nextflow.io/docs/latest/basic.html). Nextflow is a workflow manager which is also using containers to ensure efficient operation and reproducibility [9]. Nextflow is based on a succession of independent processes each having input and output. Each process can communicate with the other via channels. Snakemake (https://snakemake.readthedocs.io/en/stable/index.html) is also another workflow management system with a Python based language [10].

Programming languages, software tools and workflow languages were used to illustrate examples that could be used for bioinformatics data analysis. **Figure 1** illustrates some platforms, facilities and tools that could be used to perform bioinformatics and biostatistical analyses.

Galaxy platform

Freely accessible online platforms such as the Galaxy Community Hub (https://galaxyproject.org/) make it much easier for novice scientists and students to use bioinformatics tools. This open-source platform offers a wide range of services, from data analysis and workflow implementation to training and education for scientists. Related pages such as the Galaxy Training (https://training.galaxyproject.org/training-material/) offer structured learning environments and inspiring tutorials for setting up bioinformatics analysis step by step.

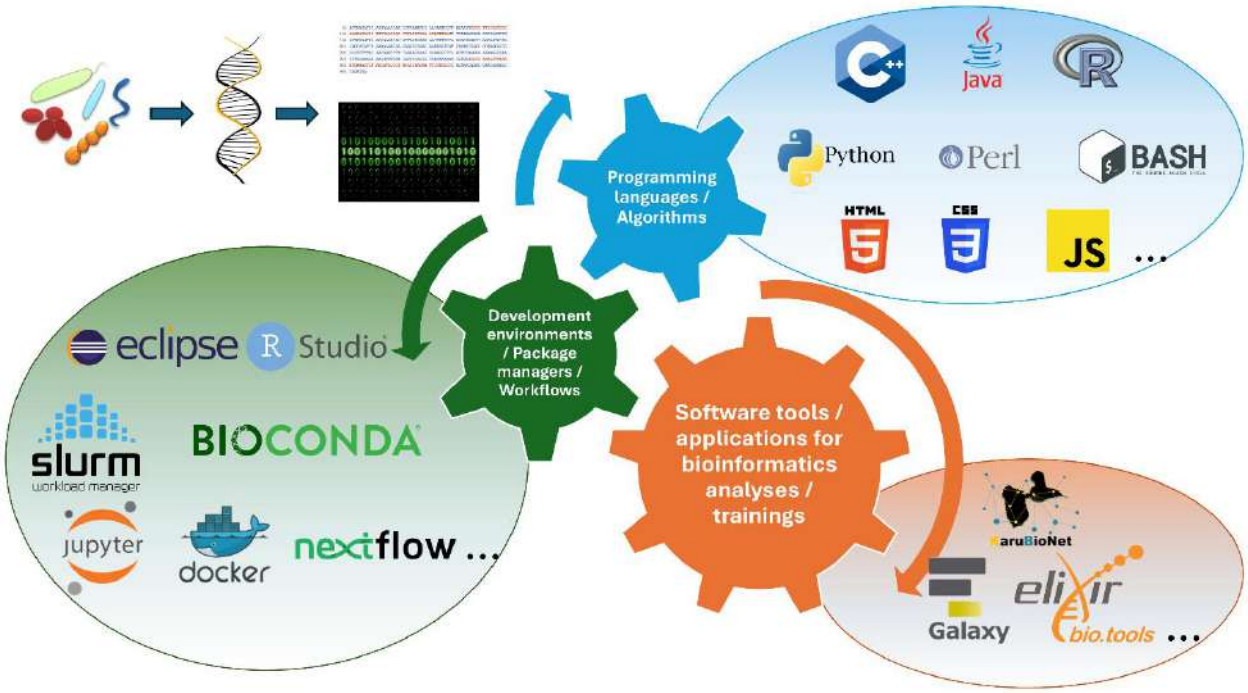


Fig. 1: *Overview of platforms, facilities and tools used for bioinformatics analysis.*

Results and Discussion

Long-read sequences analysis with Nanopore MinION

Nanopore sequencing is a third-generation sequencing approach providing long-read sequencing. It allows the sequencing of polynucleotides in the form of native DNA or RNA. This sequencing technology is widely used in many laboratories (although other long-read sequencing technologies are also available like Pacific Biosciences). The MinION (https://nanoporetech.com/products/minion) is one of the Nanopore sequencing devices and provides portable, real-time, flexible, and powerful sequencing.

The FASTQ genomic reads generated can then be processed for de novo genome assembly using tools such as Flye or Dragonflye (among others) [11]. If Illumina short-reads are also available, hybrid assembly software tools such as Unicycler [12] can be used to complement the long-reads. For deeper analyses, several dedicated bioin-formatics tools could be used from the Oxford Nanopore Technologies GitHub repository (https://github.com/nanoporetech). Basecalling and demultiplexing of the raw fast5/pod5 files can be performed directly using the MinKNOW software (https://nanoporetech.com/about-us/news/introducing-new-minknow-app). However, dedicated software tools can be used for basecalling: Guppy (https://timkahlke.github.io/LongRead_tutorials/BS_G.html) or the newest released version Dorado (https://github.com/nanoporetech/dorado) or Deepbinner (https://github.com/rrwick/Deepbinner) [13]; and for demultiplexing, the EPI2ME software tool provided by Oxford Nanopore Technologies can

be used (https://labs.epi2me.io/). Guppy can also perform the demultiplexing step in real time. Once the sequence reads have been obtained and split into each barcode, software tools dedicated to performing the quality control of the data, such as pycoQC (https://a-slide.github.io/pycoQC/) [14], are used. Note that these quality control tools have been made available in our Galaxy instance. **Figure 2** shows a simplified workflow for sequencing and processing data using the Nanopore MinION sequencer.

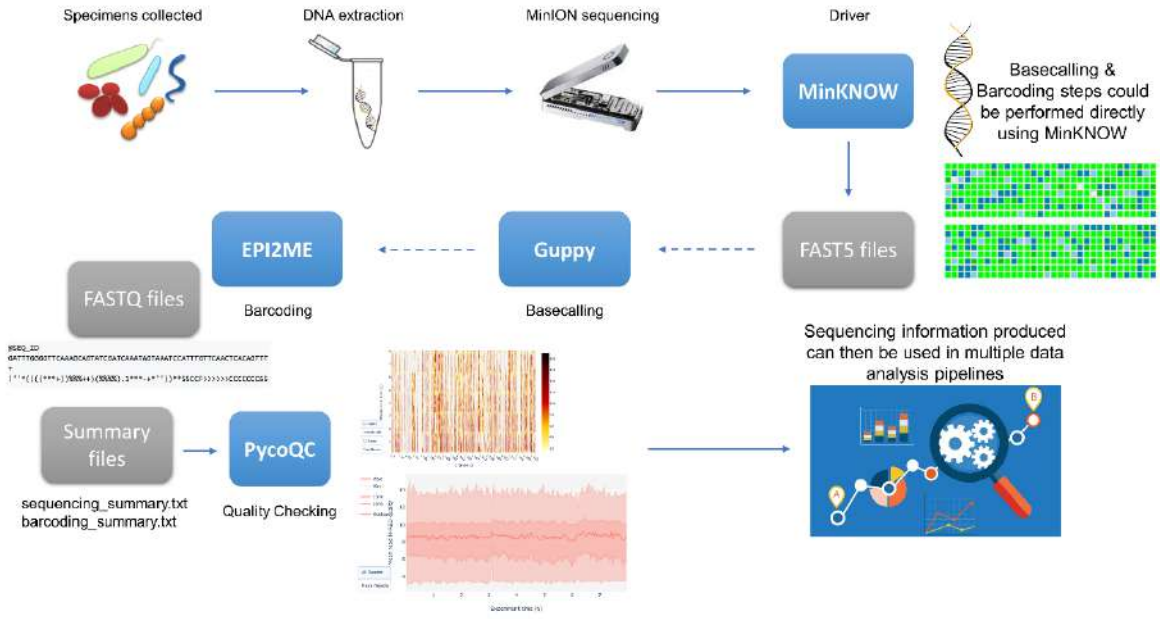
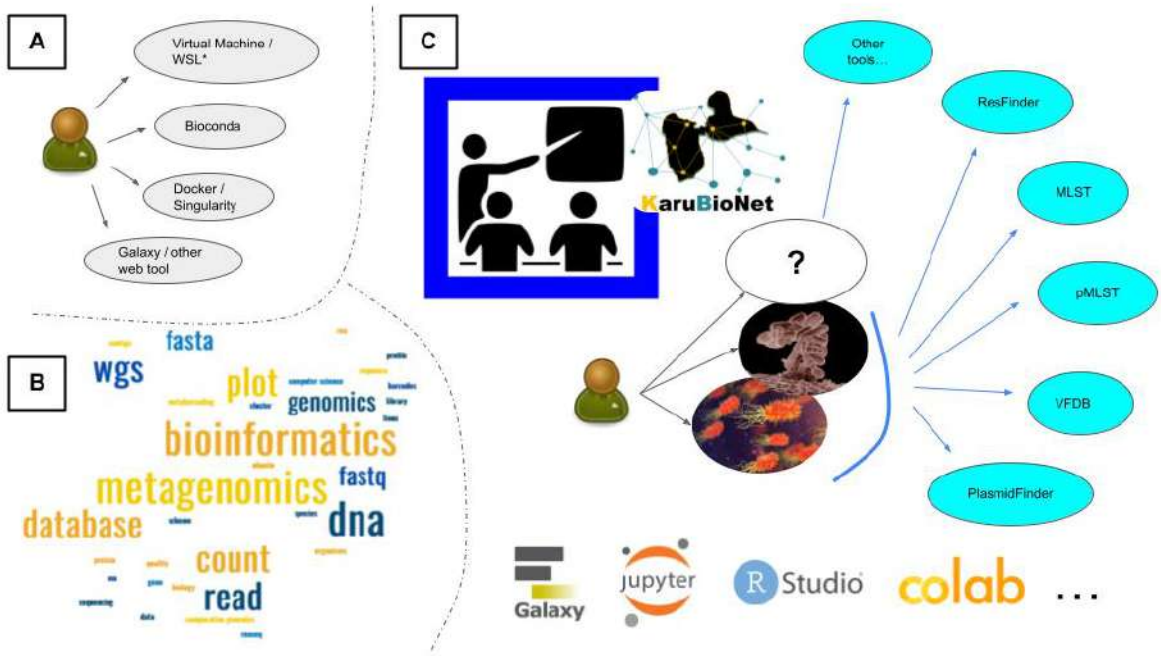


Fig. 2: *MinION sequencing and data processing workflow*.

GitHub repository for capacity building and local training

Development of specific bioinformatic training materials and software is a key step for a better understanding of real-life data surrounding us. Over the past three years, several training courses have been set up in Guadeloupe to address the analysis of massive biological data from the island's various ecosystems (marine, hospital, urban or rural, etc.), using a One Health approach. These data (generally derived from high-throughput sequencing) are often studied using prokaryotic/eukaryotic genomics, metagenomics or transcriptomics tools. **Figure 3** shows some facilities, bioinformatics and biostatistics themes and tools that could be used during training sessions.



Conclusion

In summary, we present a GitHub repository (https://github.com/karubiotools/AnssBin) designed to provide clear and accessible guides for non-bioinformatics users interested in performing their own analyzes. Our repository features a learning framework that emphasizes practical examples. We aim to continually improve this repository by adding more examples and tutorials that address the specific needs of bioinformatics and biostatistics analyses. In the meantime, we also intend to promote capacity-building in bioinformatics and growth of local bioinformatics platforms like KaruBioNet in French overseas territories (http://www.pasteur-guadeloupe.fr/karubionet.html).

References

1. Pot M, Reynaud Y, Couvin D, Dereeper A, Ferdinand S, Bastian S, Foucan T, Pommier JD, Valette M, Talarmin A, Guyomard-Rabenirina S, Breurec S. Emergence of a Novel Lineage and Wide Spread of a blaCTX-M-15/IncHI2/ST1 Plasmid among Nosocomial Enterobacter in Guadeloupe. Antibiotics (Basel). 2022 Oct 20;11(10):1443. doi: 10.3390/antibiotics11101443.
2. 2. Dereeper A, Gruel G, Pot M, Couvin D, Barbier E, Bastian S, Bambou JC, Gelu-Simeon M, Ferdinand S, Guyomard-Rabenirina S, Passet V, Martino F, Piveteau P, Reynaud Y, Rodrigues C, Roger PM, Roy X, Talarmin A, Tressieres B, Valette M, Brisse S, Breurec S. Limited Transmission of Klebsiella pneumoniae among Humans, Animals, and the Environment in a Caribbean Island, Guadeloupe (French West Indies). Microbiol Spectr. 2022 Oct 26;10(5):e0124222. doi: 10.1128/spectrum.01242-22.
3. Bankevich, A., Nurk, S., Antipov, D., Gurevich, A. A., Dvorkin, M., Kulikov, A. S., Lesin, V. M., Nikolenko, S. I., Pham, S., Prjibelski, A. D., Pyshkin, A. V., Sirotkin, A. V., Vyahhi, N., Tesler, G., Alekseyev, M. A., & Pevzner, P. A. (2012). SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. Journal of computational biology : a journal of computational molecular cell biology, 19(5), 455–477. https://doi.org/10.1089/cmb.2012.0021.
4. Seemann T. Prokka: rapid prokaryotic genome annotation. Bioinformatics. 2014 Jul 15;30(14):2068-9. doi: 10.1093/bioinformatics/btu153.
5. The Galaxy Community, The Galaxy platform for accessible, reproducible and collaborative biomedical analyses: 2022 update, Nucleic Acids Research, Volume 50, Issue W1, 5 July 2022, Pages W345–W351, https://doi.org/10.1093/nar/gkac247.
6. Grüning, B., Dale, R., Sjödin, A. et al. Bioconda: sustainable and comprehensive software distribution for the life sciences. Nat Methods 15, 475–476 (2018). https://doi.org/10.1038/s41592-018-0046-7

7. Ison J, Rapacki K, Ménager H, Kalaš M, Rydza E, Chmura P et al. Tools and data services registry: a community effort to document bioinformatics resources. Nucleic Acids Res. 2016 Jan 4;44(D1):D38-47. doi: 10.1093/nar/gkv1116.
8. Couvin D, Dereeper A, Meyer DF, Noroy C, Gaete S, Bhakkan B, Poullet N, Gaspard S, Bezault E, Marcelino I, Pruneau L, Segretier W, Stattner E, Cazenave D, Garnier M, Pot M, Tressières B, Deloumeaux J, Breurec S, Ferdinand S, Gonzalez-Rizzo S, Reynaud Y. KaruBioNet: a network and discussion group for a better collaboration and structuring of bioinformatics in Guadeloupe (French West Indies). Bioinform Adv. 2022;2(1):vbac010. https://doi.org/10.1093/bioadv/vbac010
9. Di Tommaso, P., Chatzou, M., Floden, E. W., Barja, P. P., Palumbo, E., & Notredame, C. (2017). Nextflow enables repro-ducible computational workflows. Nature Biotechnology, 35(4), 316–319. doi:10.1038/nbt.3820
10. Mölder, F., Jablonski, K.P., Letcher, B., Hall, M.B., Tomkins-Tinch, C.H., Sochat, V., Forster, J., Lee, S., Twardziok, S.O., Kanitz, A., Wilm, A., Holtgrewe, M., Rahmann, S., Nahnsen, S., Köster, J., 2021. Sustainable data analysis with Snakemake. F1000Res 10, 33. https://doi.org/10.12688/f1000research.29032.1
11. Kolmogorov M, Yuan J, Lin Y, Pevzner PA. Assembly of long, error-prone reads using repeat graphs. Nat Biotechnol. 2019 May;37(5):540-546. doi: 10.1038/s41587-019-0072-8.
12. Wick, R. R., Judd, L. M., Gorrie, C. L., and Holt, K. E. (2017). Unicycler: Resolving bacterial genome assemblies from short and long sequencing reads. PLoS Comput. Biol. 13:e1005595. doi: 10.1371/journal.pcbi.1005595
13. Wick RR, Judd LM, Holt KE (2018) Deepbinner: Demultiplexing barcoded Oxford Nanopore reads with deep convolutional neural networks. PLoS Comput Biol 14(11): e1006583. https://doi.org/10.1371/journal.pcbi.1006583
14. Leger A and Leonardi T (2019). pycoQC, interactive quality control for Oxford Nanopore Sequencing. Journal of Open Source Software, 4(34), 1236, https://doi.org/10.21105/joss.01236

Session 3: Systems Biology

Metagenome-scale metabolic modelling for the characterization of cross-feeding interactions in freshwater cyanobacteria-associated microbial communities

Juliette AUDEMARD[1], Mohamed MOUFFOK[2], Charlotte DUVAL[3], Jeanne GOT[4], Sébastien HALARY[3], Marie LEFEBVRE[2], Julie LELOUP[5], Benjamin MARIE[3], Gabriel MARKOV[6], Coralie MULLER[1], Nicolas CREUSOT[7], Binta DEME[2] and Clémence FRIOUX[1]

1 Inria, University of Bordeaux, INRAE, F-33400 Talence, France

2 Université Clermont Auvergne, INRAE, UNH, PFEM, MetaboHUB Clermont, 63000 Clermont-Ferrand, France

3 Muséum National d'Histoire Naturelle, UMR 7245 CNRS-MNHN, MCAM, Paris, France

4 Université Rennes, Inria, CNRS, IRISA, Rennes, France

5 Sorbonne Université, Univ Paris-Cité, Univ Paris-Est, CNRS, IRD, INRAE, IEES Paris, F-75005

6 Sorbonne Université, CNRS, Laboratoire de Biologie Intégrative des Modèles Marins, LBI2M, F-29680 Roscoff, France

7 INRAE, UR1454 EABX, Bordeaux Metabolome, MetaboHub, Gazinet Cestas, France

Corresponding author: juliette.audemard@inria.fr, clemence.frioux@inria.fr

Keywords

systems biology, microbial community, metabolic modelling, freshwater cyanobacteria, metabolomics

Abstract

Favoured by global changes, freshwater cyanobacterial harmful blooms (HCBs) generate increasing ecological, economical and public health challenges. Microcystis, one of the most pervasive genera of cyanobacteria, grows within a phycosphere where specialised interactions with its microbiome occur, and are suspected to influence bloom appearance and its potential toxicity.

Through metagenomic, metabolomic and metabolic modelling, we characterised twelve Microcystis phycospheres cultured after isolation from a French pond. Metagenomics revealed that associated bacteria introduce new functions to the phycosphere, while functional redundancy within and across communities remains. Metabolic reaction presence in Microcystis is consistent with their genospecies, whereas community-level metabolic landscape diverges from cyanobacteria's phylogeny. On the other hand, metabolomic results lean on metabolic output led by cyanobacteria. Metabolic modelling and identification of toxic secondary metabolites biosynthetic gene clusters further highlighted differences between phycosphere metabolic capabilities and the importance of manual curation of secondary metabolism in metabolic networks. These findings deepen understanding of Microcystis' phycosphere functioning, demonstrate the relevance of multi-omics systems biology approaches, and lay the ground for further characterisation of freshwater HCB's microbial interactions and inter-species complementarity.

Met4J: a library, a toolbox and a workflow suite for graph-based analysis of metabolic networks

Ludovic COTTRET[1,2], Marion LIOTIER[1,2], Louison FRESNAIS[1], Meije MATHE[1], Fabien JOURDAN[1,2] and Clément FRAINAY[1,2]

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

2 MetaboHUB, National Infrastructure of Metabolomics and Fluxomics, Toulouse, France

Corresponding author: clement.frainay@inrae.fr

Keywords

Metabolic Network, Graph Theory, Metabolism, Workflow , Galaxy

Abstract

Graph algorithms are essential tools for network analysis in various domains, including biology. Despite successful applications to metabolic networks, including several developments specific to these models, few implementations are openly available. Furthermore, the exchange format adopted for most genome-scale models is incompatible with the main generic graph-analysis libraries. We present Met4J, an open-source library dedicated to the structural analysis of metabolic models and their manipulation, as well as a toolbox encompassing implementations of analyses relevant to metabolism-related research. We exemplify the potential of Met4J by creating a workflow for the construction and analysis of a holobiont network. Met4J's source code, executable JAR and containers are available at https://forgemia.inra.fr/metexplore/met4j and the library artifact is accessible through the Maven central repository. High-level applications are also available on a Galaxy interface.

Regulatory response of maize to water deficit mediated by distal *cis*-regulatory elements

Thomas-Sylvestre MICHAU[1], Tristan MARY-HUARD[1], Maud FAGNY[1]

1 GQE-Le Moulon, 12 rue 128, 91190 Gif-sur-Yvette, France

Corresponding Author: thomas.michau@inrae.fr

Keywords

Response to water deficit, Maize genomic, Gene regulatory network, *Cis*-regulatory elements

Abstract

Climate change is intensifying summer droughts in Europe, significantly affecting plant growth and crop yields.The flowering of maize happens during this high-risk period, and the water deficit stress results in kernel abortion among other consequences, seriously impacting yield. Understanding the genomic basis of maize responses to water deficit is therefore crucial for agricultural adaptation. Plants regulate gene expression in response to environmental changes through signaling pathways, where transcription factors (TFs) interact with cis-regulatory elements (CREs). Two main categories of CREs exist: (i) proximal CREs, or promoters, and (ii) distal CREs (dCREs), which include enhancers and silencers. dCREs can regulate multiple genes across various cell types and influence gene expression in a complex, context-dependent manner. While promoters have been extensively studied, dCREs remain underexplored in maize due to genome assembly challenges and the absence of specific epigenetic markers. However, recent technological advances, including long-read sequencing and the discovery that most maize dCREs are unmethylated, facilitate their identification and functional characterization.

Traditional gene regulatory network (GRN) analyses rely on co-expression networks, which assume that TF effects are directly correlated with their expression levels. To address this limitation, we propose a co-regulation approach that considers the physical interaction potential of TFs with CREs and the correlation between the expression of target genes regulated by the same TF.

Our study focuses on identifying dCREs involved in maize responses to water deficit in the reference inbred line B73. We constructed an initial regulatory network using methylation data and genome annotation, which was refined using message-passing-based inference to generate tissue- and condition-specific networks. Comparative analysis revealed differential gene regulation across tissues, particularly in leaves, where genes associated with photosynthesis and stress responses were significantly affected. These findings highlight the utility of CRE-based networks for identifying key regulatory elements in maize drought responses.

Introduction

Climate change is increasing the risk of summer droughts in Europe, affecting plant growth and development. The flowering of maize happens during this high-risk period, and the water deficit stress results in kernel abortion among other consequences, seriously impacting yield [1]. Moreover, different

maize lines exhibit various degree of tolerance to a similar water deficit, a trait that is polygenic, and often inversely correlated to yield performance. This suggests that there is still space for water deficit tolerance improvement in elite maize lines used to produce the cultivated maize hybrids. As the second most cultivated crop in Europe, understanding the complex mechanisms governing environmental stress response of maize and their genomic bases is thus of agricultural and economical interest.

Plants respond to their environment thanks to signaling pathways that alter the regulation of gene transcription, ultimately affecting plant phenotypes. Last effectors of the response, transcription factors (TFs) bind to *cis*-regulatory elements (CREs) and interact with the transcription machinery to modulate target genes expression level. *Cis*-regulatory elements are thus good candidate to explain differences in response to water deficit. In plants, two main categories of *cis*-regulatory elements (CREs) can be distinguished: (i) proximal CREs, which correspond to promoters, and (ii) distal CREs (dCREs), located more than 2 kb from the transcription initiation site, that encompass both the enhancers and the silencers, which respectively increase or decrease the transcriptional activity of their target genes. A single dCRE can regulate different genes in different cell types. Conversely, a gene can be regulated by multiple dCREs [2]. In maize, it is estimated that 34% of dCREs potentially regulate multiple genes simultaneously, and 25% to 40% do not target the nearest gene [3]. The interactions between TFs and genes through CRE are thus the basis of a complex network that can be rewired in response to environmental cues such as water deficit.

While promoters are well studied for their role in gene expression regulation, dCREs play a major role in controlling plant development [2] and environmental responses by regulating gene expression over time in a cell-type-specific manner [4]. However, dCREs remain an unexplored component of transcriptional regulation in maize response to environmental factors, due to both the difficulty to assemble the maize repeats-rich intergenic regions and the lack of specific epigenetic marks. Advances in long-read technologies, and the recent discovery that most maize dCRE are unmethylated [6] opened up a path to identify them.

Assessing the effect of TFs in gene regulatory networks and their role in response to environmental cues remains mostly achieved by building co-expression networks. While being accessible, requiring only gene expression data, this method presents limits. Indeed, because it focuses on expression data, it relies strongly on the hypothesis that the effects of TFs are correlated with their transcription level, which has been invalidated [9,10]. In this study we propose a co-regulation approach which is not based on the correlation between regulator and target but on the possibility of regulator to bind a CRE nearby its potential targets, and on the correlation between the expression of the different targets of a same regulator.

Here, we present an approach aimed at identifying dCREs involved in the regulation of gene expression under water deficit conditions in the maize reference inbred line B73. A prior was built using methylation data and genome annotation. This network was then refined through message-passing-based inference to

generate tissue- and condition-specific networks. By comparing condition- and tissue-specific networks, we identified genes and functions whose regulation was affected by watering conditions.

Materials and Methods

Biological material

Expression data

In this study we used genomic, transcriptomic and epigenetic data from the maize reference line B73, part of the dent genetic group, from the Corn Belt (United States). The samples used for the mRNA data correspond to five tissues: internodes, ears, mature leaves, panicles, and silks; under two water conditions: no stress (well-watered, WW) and water defict (WD). Three biological replicates were sequenced for every tissues and conditions, totaling 30 samples. All the tissues were sampled at the same developmental stage, at pollen shed, corresponding to the flowering of the tassel, the male flower.

Tissues were sampled and RNA was extracted and sequenced according to the protocol described in Fagny et al., 2021 [3]. Also the RNA-seq data were pre-processed and aligned following the method described in Fagny et al., 2021 [3], but using the Zm-B73-REFERENCE-NAM-5.0 assembly of the B73 maize genome.

The expression data were normalized using the SNAIL method [5]. This quantile-based normalization approach operates in groups, allowing the normalization of a dataset containing samples from different tissues while preserving the tissue-specific gene expression distributions.

Methylation data

To identify low methylation region, we used already published analyse pipeline from Crisp et al., PNAS, 2020[6], on our own bisulfite sequencing data set. Bisulfite sequencing convert unmethylated cytosines into uracil. After replication, the uracils are replaced by thymines, while only methylated cytosines are preserved in the sequences [7]. After replication, the uracils are replaced by thymines, while only methylated cytosines are preserved in the sequences. This alteration allow to identify unmethylated region (UMRs) by mapping the reads using bsmap[8]. The predicted dCREs are low-methylation intergenic sequences, so genic and promotor regions were removed from the dataset using the Zm-B73-REFERENCE-NAM-5.0 assembly.

Transcription factors data

The PlantTFDB website (https://planttfdb.gao-lab.org) is a database that catalogs transcription factors (TFs) for various plant species, including maize, along with their binding motifs. For maize, 259 TFs are listed. In addition to this data, binding motifs from Jaspar (https://jaspar.elixir.no), are also included, totaling 98 motifs for maize. Among these, 8 were already present in the PlantTFDB database. The detection of TFBS was done using FIMO from the MEME suite [11] on promotor and UMR sequences. The detected sites were considered significant if their Benjamini-Hochberg adjusted *p*-value was below 0.05 (5% false discovery rate). This TFBS detection was performed on promotors and dCRE separately.

Network inference

The NetZooPy package [12] is used to infer and analyse gene regulatory networks. The package notably includes the PANDA algorithm [13], which generates a weighted graph based on three types of data: a gene

co-expression matrix, a protein-protein interaction matrix between transcription factors (TFs), and a prior listing all possible TF-gene regulations relationships. The result represents the regulatory relationships between TFs and genes at an organism scale, across all considered samples. To achieve this, PANDA initializes the regulatory network based on the prior, which corresponds to all possible interactions in the maize lineage. The edge weights of this initial network are then updated using a message-passing procedure.

The LIONESS algorithm [14] produces tissue-condition specific networks through a linear combination of two PANDA networks: the complete network, with all samples, and a network excluding a sample corresponding. This process is repeated for each sample, resulting in a final matrix where each edge is assigned a weight for each sample. The weights can either be positive or negative, as they represent the contribution of the sample to the complete network. Positive edges thus show greater regulation by transcription factors on genes in the sample than in the complete network, while negative edges show lower regulation.

Prior

We built the prior network using all CRE. Promoter's target genes were immediately deduced from the genome annotation. The non-genic UMRs that contained at least one predicted TFBS were considered as potential dCREs, and their candidate target genes were all the genes being located at less than 100 kb. This threshold was chosen based on the results of Lu et al. 2019 Nature Plants [15], that shows that most dCRE are located at about 60kb from transcription initiation site in maize.

Differential targeting analysis

Differential targeting analysis was performed using a generalized linear regression model and implemented using the R Bioconductor limma package [16]. In this analysis, for each gene in the network, we defined the indegree as the sum of the edges in the regulatory network targeting the gene, and we compared them between the two watering conditions. The genes were considered differentially regulated if their Benjamini-Hochberg adjusted p-value was below 0.05 (5% false discovery rate). The Fold Change (FC), usually generated in differential analysis, is here used as indicative of the differences in regulations: logFC> 0 were classified as more regulated in WD, and logFC < 0 were classified as more regulated in WW.

Those analysis are associated with a *t*-statistic, based the logFC and its standard error. This statistic is used as a rank for genes in the fgsea analysis for gene ontology enrichment.

Gene ontology enrichement

Gene ontology enrichment analyses were performed using the FGSEA (*Fast Gene Set Enrichment Analysis*) [17] algorithm implemented in the fgsea R package. Genes were ranked based on their *t*-statistic. The functions were considered differentially regulated if their Benjamini-Hochberg adjusted *p*-value was below 0.05 (5% false discovery rate).

Results

Prior and data description

In order to infer tissue- and condition-specific networks, we first built a prior regulatory network that combines regulation from both types of CRE without distinguishing the source of the TFBSs (see Material and Methods). It includes a total of 37,277 genes and 35,185 UMRs, for around 70k regulatory sequences and 79 different motifs (Tab. 1). It leads to a total of 206302 edges between TFBSs and genes. The figure showsWe observe a great variation in the number of TF targeting each gene, with an average of 23 TFs targeting a given gene and a median at 26 TF (Fig. 1).

Conversely, some TFs were more connected in the prior network than others (Fig. 2). The proportion of CRE containing the different TF shows great variations from $1.4x10^{-3}$ % to 50.5 % (Fig. 3). Comparing these proportions to the number of targeted genes, The 4 TFs presenting the highest number of regulations are in the 6 most abundant TFs in the network.

Expression data analysis

With the previously described prior, expressions data are second type of data used for the network inference. We obtained normalized gene counts for five tissue in two watering conditions and three replicates (see Material and Methods). We first visualized these data performing a principal component analysis (PCA) and a correlation matrix analysis.

The figure shows a PCA of the expression data, with each dot corresponding to a different sample. The PCA clusters expression profiles by tissue. This global analysis reveals a tissue-specific organization of gene expression. The correlation matrix of the expression data shows a higher correlation between samples from the same tissue.

Analysis of B73 network inferences

Global analysis of tissue-specific networks

We obtained sample-specific regulatory networks using PANDA to infer a global summary network, and LIONESS to infer sample-specific networks, both from the NetZoo suite. The edges of these tissue-specific networks can have either negative or positive weights. In the context of a specific edge in a specific sample, a negative weight indicates a reduction in regulation compared to the global network, while a positive weight indicates an increase. To ensure that the inferred networks accurately represent the specific response of the different combinations of tissues and conditions, we performed preliminary analyses. These included PCA and Spearman correlation analyses of sample-specific network edges.

Fig. 5A shows that the PCA clusters networks by tissue. This global analysis of the networks highlights a tissue-specific organization of regulation. The correlation matrix of the networks shows a higher correlation between samples from leaves, silks, and tassels, each forming distinct groups. Moreover, for leaves, both the PCA and the correlation matrix distinguish the two watering conditions, separating them in the PCA and associating them with lower correlation values.

Because we were particularly interested in the role of dCRE, we extracted the edges associated with them from the sample-specific networks, without the promoters. The PCA and correltion matrix restricted to the dCRE regulations show similar results of those of the global network.

From expression to regulation profiles, leaves, silks, and tassels maintain their tissue specificity, as visualized in the PCAs and correlation matrices. However, for ear and internode networks, this pattern is less clear, with only slightly higher correlations within the same tissue compared to others.

Differential targeting between WW and WD by tissue

The significance of regulatory changes was assessed through a differential analysis of edge weights between both watering conditions. Although a large number of edges changed between the two conditions, only 7% were significant in the leaf, and 0,001% in the tassel. These tissues shared 3 significant edges. For other tissues no edges presented significant variations. Considering this result, the next sections will focus exclusively on leaf data.

The differential targeting analysis of regulatory relationships showed that only a fraction of regulations is significant to each tissue (Fig 6A). The scatterplot in Fig 6B compares edge weights between both watering conditions. Each dot represents an edge. Dots in red or blue correspond to significant regulatory variations of a TF on a gene depending on water conditions. Red dots indicate an increase in regulation under water deficit conditions, while blue dots indicate a reduction. To identify functions affected by drought, we examined genes and their targeting scores, summing all regulatory edge weights per gene (indegree). As in LIONESS networks, a negative value indicates a reduced regulation, positive values an increase. Indegree values ranged from -1040 to 2288 across all networks and from -3017 to 6252 for dCRE-only networks. Comparing these scores between watering conditions with limma identified 4k genes with significant targeting variations in leaves. The differential analysis identified 4023 genes differentialy regulated in leaves, and 16 tassels. Those tissues shared 1 significant gene. No differentially regulated gene was identified for the other tissues.

The Fig 7 compare the difference in expression and regulation using the water deficit condition as refrence. When the difference in regulation is positive, it show more regulation in the WD condition, (same with expression level).Increased targeting in WD is thus associated with a decreased expression, and decreased targeting with an increased expression. For both dCRE-only and CRE regulations, the regulation differences primarily indicate inhibition mechanisms.

A gene ontology enrichment of the leaves results analysis, using fgsea, identified 28 enriched terms including functions linked to photosynthesis, a function known to be altered by water deficit: photosynthesis (GO:0015979, p-value = $5{,}6\text{x}10^{-10}$), light reaction (GO:0019684, p-value = $7{,}9\text{x}10^{-8}$). Responses to environmental factors were also enriched, including abiotic response (GO:0009628, p-value = $2{,}0\text{x}10^{-5}$), and response to temperature stimulus (GO:0009266, p-value = $1{,}1\text{x}10^{-2}$). Endogenous responses like response to hormone (GO:0009725, p-value = $2{,}3\text{x}10^{-2}$) were also enriched. A similar analysis based

solely on dCRE regulations showed enrichments for 6 terms, including those related to photosynthesis and response to abiotic stress but not the ones related to endogenous response.

Discussion

This study, based on an analysis of gene regulatory networks derived from CRE regulation, aimed to capture differences in gene regulation in response to water deficit across various maize tissues. The inferred regulatory networks are tissue-specific, showing high correlation within tissue groups, particularly in silks, tassels, and leaves. The latter two exhibit differentially regulated genes, particularly in leaves. Focusing on the leaves, the relevance of the analysis is underlined by the gene ontology enrichment that revealed terms associated with photosynthesis and stress response. These findings highlight the relevance of using CRE-based networks to identify target genes involved in water deficit response, showing coherent regulations .

Because the prior network is based on the possibility of interaction of TF with target genes, it is highly dependent of the detection of these TFs. Some of them present a higher number of edges that could be attributed to the abundance of certain TFBS motifs compared to others. This abundance could be linked to a more frequent identification of the specific motif in sequence by FIMO, or a higher abundance of this motif in UMRs. In the last case, if a TF interacts with an higher number of dCREs than promotors, its number of interactions is mechanically increased in the global network because of the higher number of candidate target genes for each dCRE.

In the ongoing process of reconstructing dCRE-specific regulations, identifying a relevant set of edges is crucial. Identifying differential regulation thus helps minimize the risk of including TFBS that were detected with high confidence but are not relevant in the regulatory context. Nonetheless restraining the analysis to dCRE regulation only reduced the number of enrichment or differential regulation. This could be explained by the reduction of number of TF considered when focusing on dCRE only. This result shows the need to adapt the regulatory score to be relevant when extracting regulation for those regulators.

Finally, the analysed regulation profiles are specific of a maize line, and our conclusions cannot be extended across different lines. Similar analyses will thus be conducted on other maize lines to identify their specific regulation profiles. The addition of lines from various genetic pools to the analysis will improve our knowledge of gene expression regulation in response to water deficit. This will also allow us to link any variation in the regulatory network structure to genetic variations in regulatory sequences, but also to maize phenotypic response to water deficit.

Figures and tables

	Promotors	dCRE
Number of items	35276	35185
Number of regulation (CRE-gene)	35276	170996
Number of motifs	71	59

Total number of TFBS	308711	411092

Tab 1 : Description of the global caracteristic of the CRE depending of their type

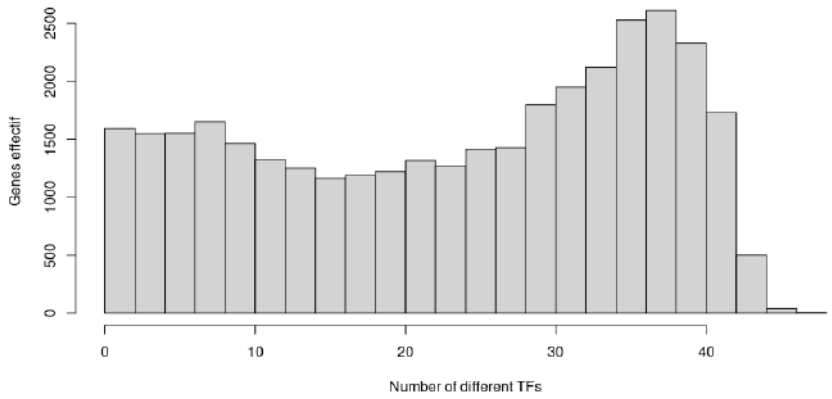


Fig 1 : Distribution of the number of genes depending of the number of targeting TF

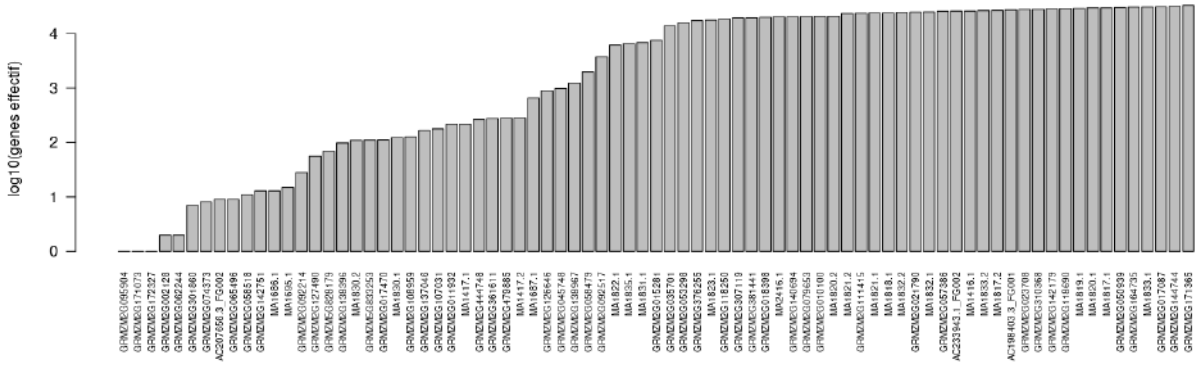


Fig 2 : Distribution of the number of genes targeted by the different TF

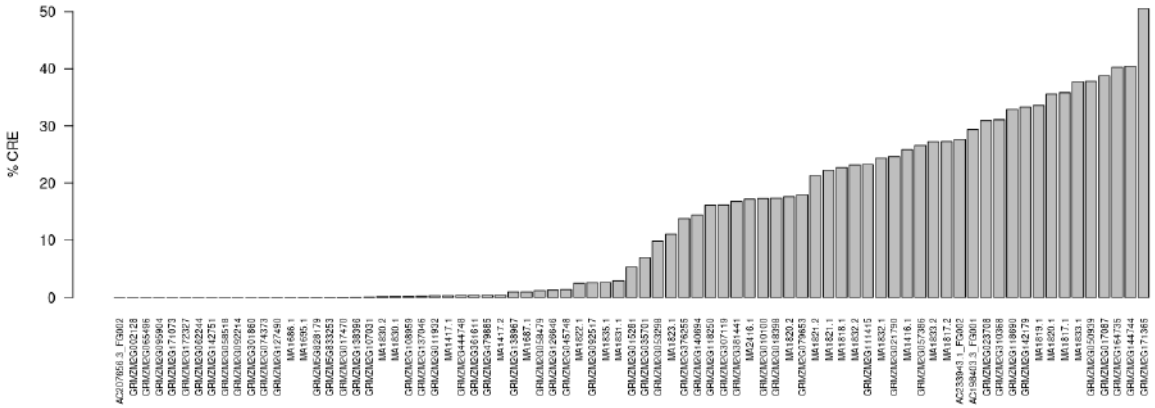


Fig 3 : Distribution of the proportion of CRE containing the different TFs

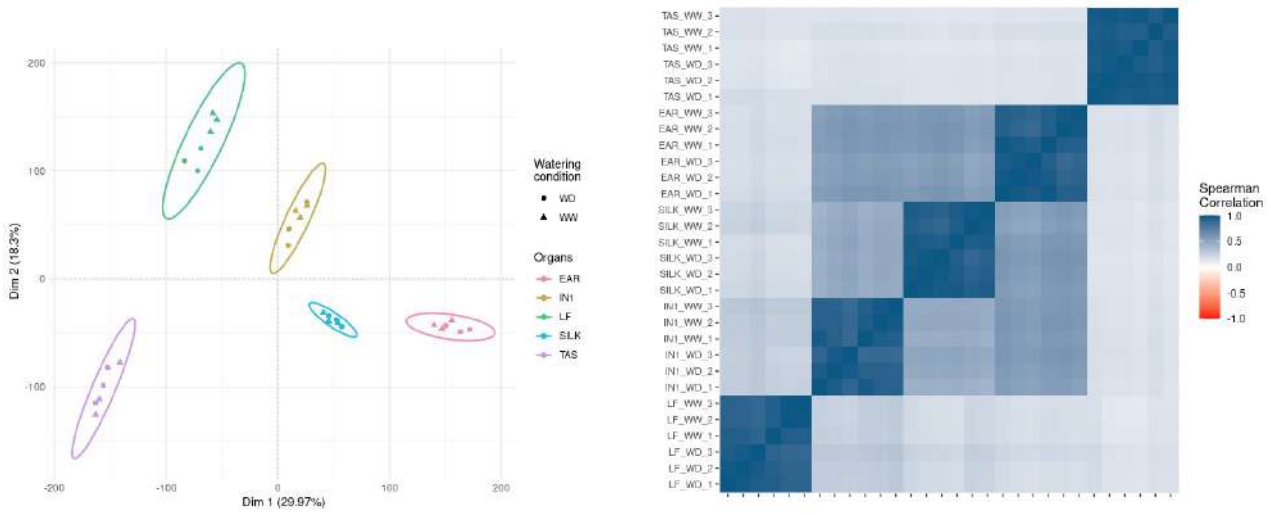


Fig 4 : Descriptive analysis of the B73 expression data : PCA (A), correlation matrix (B)

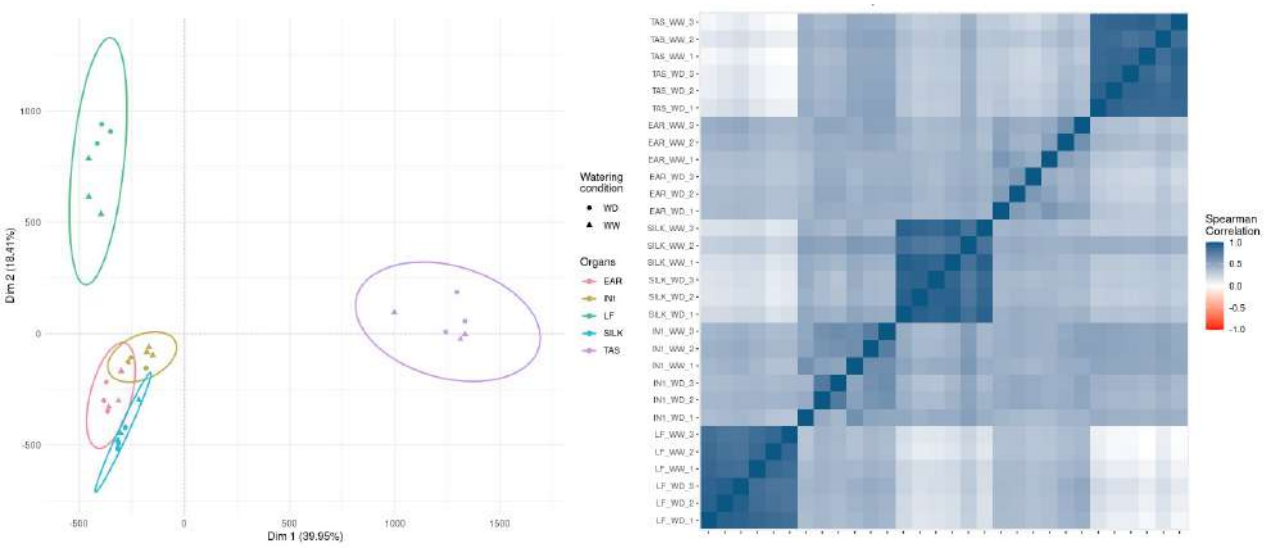


Fig 5 : Descriptive analysis of the NetZoo networks : PCA (A), correlation matrix (B)

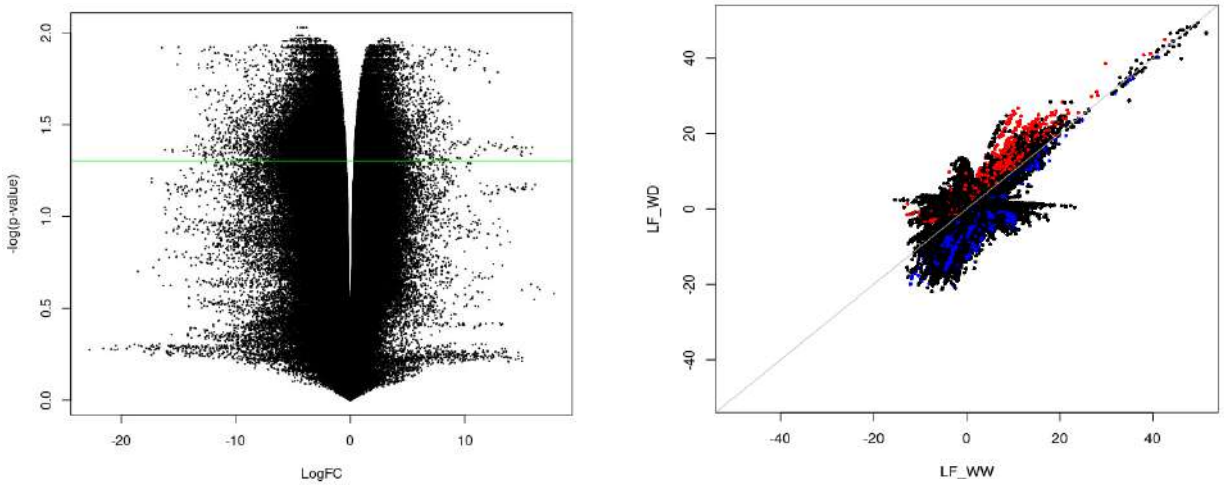


Fig 6 : Differential analysis of edges weight of leafs, depending of water condition

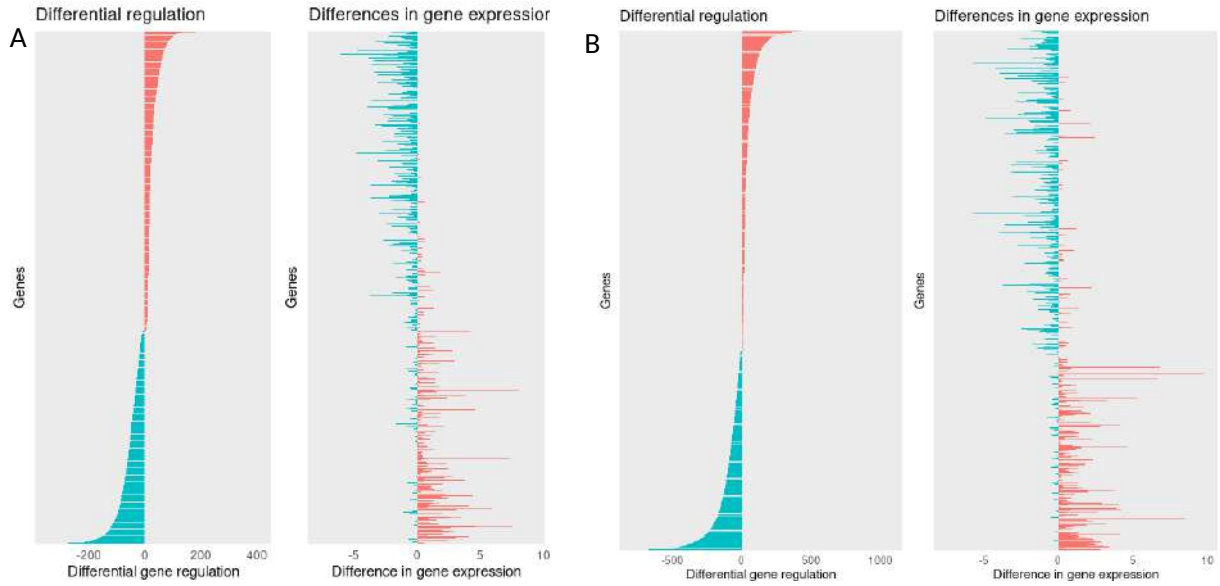


Fig 7 : Comparaison between differential regulation and gene expression for CRE (A) and dCRE (B) networks of leafs

Script availability

Analysis and figures : https://forgemia.inra.fr/thomas.michau1/maize_grn_analysis_jobim2025

NetZoo : https://netzoo.github.io/

References

1. NeSmith DS, Ritchie JT. Effects of soil water-deficits during tassel emergence on development and yield component of maize (Zea mays). Field Crops Research. 1992 Jan;28(3):251–6.

2. Ricci WA. Widespread long-range cis-regulatory elements in the maize genome. Nature plants. 2019 Dec;5:1237–49.

3. Fagny M, Kuijjer ML, Stam M, Joets J, Turc O, Rozière J, et al. Identification of Key Tissue-Specific, Biological Processes by Integrating Enhancer Information in Maize Gene Regulatory Networks. Front Genet. 2021 Jan 11;11:606285.

4. Spitz F, Furlong EEM. Transcription factors: from enhancer binding to developmental control. Nat Rev Genet. 2012 Sep;13(9):613–26.

5. Hsieh PH, Lopes-Ramos CM, Zucknick M, Sandve GK, Glass K, Kuijjer ML. Adjustment of spurious correlations in co-expression measurements from RNA-Sequencing data. Boeva V, editor. Bioinformatics. 2023 Oct 3;39(10):btad610.

6. Crisp PA. Stable unmethylated DNA demarcates expressed genes and their cis-regulatory space in plant genomes. PNAS. 2020 Sep 22;117(38):23991–4000.

7. Darst RP, Pardo CE, Ai L, Brown KD, Kladde MP. Bisulfite Sequencing of DNA. CP Molecular Biology [Internet]. 2010 Jul [cited 2025 Mar 31];91(1). Available from: https://currentprotocols.onlinelibrary.wiley.com/doi/10.1002/0471142727.mb0709s91

8. Xi Y, Li W. BSMAP: whole genome bisulfite sequence MAPping program. BMC Bioinformatics. 2009 Dec;10(1):232.1.

9. Zaborowski AB, Walther D. Determinants of correlated expression of transcription factors and their target genes. Nucleic Acids Research. 2020 Nov 18;48(20):11347–69.1.

10. Brewster RC, Weinert FM, Garcia HG, Song D, Rydenfelt M, Phillips R. The Transcription Factor Titration Effect Dictates Level of Gene Expression. Cell. 2014 Mar;156(6):1312–23.

11. Grant CE, Bailey TL, Noble WS. FIMO: scanning for occurrences of a given motif. Bioinformatics. 2011 Apr 1;27(7):1017–8.

12. Guebila MB, Wang T, Lopes-Ramos CM, Fanfani V, Weighill D, Burkholz R, et al. The Network Zoo: a multilingual package for the inference and analysis of biological networks [Internet]. Genomics; 2022 May [cited 2022 Aug 9]. Available from: http://biorxiv.org/lookup/doi/10.1101/2022.05.30.494077

13. Glass K, Huttenhower C, Quackenbush J, Yuan GC. Passing Messages between Biological Networks to Refine Predicted Interactions. Semsey S, editor. PLoS ONE. 2013 May 31;8(5):e64832.

14. Kuijjer ML, Tung MG, Yuan G, Quackenbush J, Glass K. Estimating Sample-Specific Regulatory Networks. iScience. 2019 Apr;14:226–40.

15. Lu Z, Marand AP, Ricci WA, Ethridge CL, Zhang X, Schmitz RJ. The prevalence, evolution and chromatin signatures of plant regulatory elements. Nat Plants. 2019 Dec;5(12):1250–9.

16. Ritchie ME, Phipson B, Wu D, Hu Y, Law CW, Shi W, et al. limma powers differential expression analyses for RNA-sequencing and microarray studies. Nucleic Acids Research. 2015 Apr 20;43(7):e47–e47.

17. Korotkevich G, Sukhov V, Budin N, Shpak B, Artyomov MN, Sergushichev A. Fast gene set enrichment analysis [Internet]. 2016 [cited 2025 Mar 31]. Available from: http://biorxiv.org/lookup/doi/10.1101/060012

Methods for a species-specific genome-scale metabolic model designed for eukaryotes and applied to the *Ascophyllum nodosum macroalga*

Pauline HAMON-GIRAUD[1], Benoît BERGK PINTO[2], Jeanne GOT[1], Coralie ROUSSEAU[2], François THOMAS[2], Erwan CORRE[3], Simon DITTAMI[2], Gabriel MARKOV[2] and Anne SIEGEL[1]

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

[2] Sorbonne Université, CNRS, Laboratoire de Biologie Intégrative des Modèles Marins, LBI2M, F-29680 Roscoff, France

[3] Sorbonne Université, CNRS, ABiMS bioinformatic platform, F-29680 Roscoff, France

Corresponding author: pauline.hamon-giraud@irisa.fr

Keywords Genome-scale metabolic models reconstruction, Biomass function, Pathway analysis, Microbiota, *Ascophyllum nodosum*

Abstract *Genome-scale metabolic models (GEMs) are essential tools for studying metabolism, either for comparative analyses or to investigate interactions between organisms. However, genome annotation, biomass formulation, and network gap-filling are key steps in constructing a relevant GEM and ensuring the biosynthesis of specialized metabolites. We present a pipeline to integrate extensive biological knowledge (genomes of closely related species, metabolic profiling studies, potential interactions with microbiota) about an eukaryotic organism in order to generate high quality GEMs. To manage genome annotation limitations, the pipeline relies on a GEM reconstruction tool that propagates annotations across closely related species through the identification of orthologous genes. It also pays particular attention to biomass formulation, using a set of metabolomic studies to create a consensus biomass composition that seeks to closely reflect biological reality, such as incorporating specialized metabolites and their precursors. The gap-filling stage of the pipeline uses a semi-automated curation process for added reactions, taking into account the presence of orthologous genes, occurrence in phylogenetically related species and potential interactions with the organism's microbiota. The final GEM applied to the brown alga Ascophyllum nodosum comprises 3,536 metabolites and 3,072 biochemical reactions, predicting the synthesis of 1,023 compounds from 38 seawater-derived metabolites. Almost all reactions (99.98%) are linked to an enzyme supported in the algal genome. This refined model provides a framework for studying host-microbiota metabolic complementarity. This pipeline offers a scalable and robust method for reconstructing high-quality GEMs in emerging eukaryotic model organisms, improving metabolic network accuracy and expanding our understanding of species-specific metabolism. It also sheds lights on the various level of knowledge related to the synthesis pathways of the biomass, paving the way to future studies to be undergone.*

Introduction

Genome-scale metabolic models (GEMs) are powerful tools used in various metabolic studies, including predicting the production of target metabolites, assessing the metabolic variability of an organism under various conditions, identifying enzyme functions, and modeling interactions between multiple cells or organisms [1]. However, their automatic reconstruction faces challenges due to various sources of uncertainty, such as the accuracy of genome annotation, biomass formulation, and the network gap filling procedure [2]. Moreover, taking into account the biosynthesis of specialized metabolites is a key challenge in improving the quality of a species-specific GEM [3]. To address those limitations, we propose a semi-automatic pipeline designed to optimize GEMs reconstruction of an eukaryotic organism by integrating as much as possible the available biological knowledge from several sources : reactions databases, metabolomic data or literature-based biomass data, annotated genomes from phylogenetically close species and associated microbiome MAGs or metabarcoding data. This approach emphasizes explainability, ensuring that all elements integrated into the network can be traced back to their sources, enabling an assessment of their relevance.

This pipeline was developed within the study of the macroalga *Ascophyllum nodosum*, a brown alga abundant along the coasts of Brittany and of industrial interest due to its biostimulant properties [4]. Our method to reconstruct the GEM first involved the use of AuCoMe [5], a bioinformatics tool that propagated functional annotations across a corpus of 62 closely related brown algae and outgroup of other stramenopiles species[6]. Next, we identified a set of 75 weighted compounds supported by a set of literature studies that were ponderated to define a generic algal biomass. To ensure network functionality, we implemented both manual and automated curation procedures, leveraging a discrete dynamical framework gap-filling tool (Meneco) and integrating data from both the associated microbiota [7] and closely related brown algae [6].

The final network resulting from the pipeline includes 3,536 metabolites and 3,072 biochemical reactions. It predicts the synthesis of 1,023 compounds by metabolic pathways initiated from 38 metabolites reflecting seawater composition, almost each (99.98%) reaction catalyzed by an enzyme supported by a corresponding genetic sequence in the genome. All 75 biomass compounds have a possible biosynthesis pathway described. The resulting GEM includes specific features that enable studies on metabolic complementarity between the algal host and its associated microbiota while minimizing false positives related to reactions potentially originating from microbial catalytic activity.

Methods

Genomic and metagenomic data. The method used WGS data related to *A. nodosum* obtained in the Phaeoexplorer [6] project, sequenced with both Illumina and Nanopore technologies. To improve annotations, additional data from 61 stramenopile genomes, including 45 brown algae, were integrated using AuCoMe v0.5.1 [5], an automated tool designed to build GEMs while handling annotation heterogeneity. To predict metabolic pathway completions in *A. nodosum*, potentially arising from holobiont complementarity, microbiota data were integrated from various sources. These includes bacterial and fungal 16S metabarcoding data [7], fungal LSU and SSU data [8], and bacterial and fungal WGS datasets [9,10,11,12].

GEMs reconstruction tools. We used AuCoMe [5] v0.5.1 to enrich poorly annotated species by leveraging expertly annotated model species. Computations were performed on the GenOuest cluster (https://www.genouest.org/) (22 CPUs, 200GB RAM) using "filtering" option for orthology. To interpret metabarcoding and taxonomic data into GEMs, the EsMeCaTa [13] v0.2.12 tool (Estimating Metabolic Capabilities from Taxonomic Affiliations) was employed to infer functional annotations and protein sequences based on Uniprot knowledge. The microbiota GEMs were then produced using a command-line parallelized version PathwayTools (Mpwt python package) [14] from EsMeCaTa outputs and WGS data.

Discrete-based simulation of the production of biomass compounds and gap-filling. Simulations of producible metabolites, i.e., the metabolic modeling, were performed with Mene tools v3.3.0 [15]. The tool requires a list of available nutrient compounds, referred to as seeds, which initialize the inference of other reachable, i.e., producible, metabolites in the network. This step is referred to as network expansion, the formalism used in the dynamical system is the one of metabolic modeling associated with a Boolean semantic. It is used in the Meneco tool [16], which, in the genome-resolved approach, takes into account the added-value brought by a set of additional reactions, therefore suggesting the producibility of new metabolites resulting from adding reactions to the GEM.

In our case study, we used the full Metacyc database as a set of reactions. This set was furthermore enriched with artificial reactions linking all compounds to their higher-level (more generic) compounds class. The 26,866 reactions added with this approach made it possible to connect pathways described with various levels of precision in the Metacyc database.

Results

A pipeline for the construction and curation of eukaryotic genome-scale metabolic network from multiple eukaryotic genomes and associated microbiomes. The pipeline starts with AuCoMe, which generates draft GEMs with PathwayTools [17] v26.0 which are then refined through orthologous gene propagation and metabolic pathway completion with spontaneous reactions. The final dataset comprises 62 GEMs, averaging 2,986 reactions (2,902 for *A. nodosum*), with a standard deviation of 335. In contrast, the draft GEMs before orthology-based enrichment contained an average of 2,200 reactions, with a standard deviation of 508 (Fig. 1A).

For the microbiota, a total of 449 GEMs were reconstructed. From unique taxonomic assignments of metabarcoding data [7], 316 bacterial and 13 fungal networks were generated. An additional 15 fungal networks were obtained from taxonomic assignments from endophitic fungi isolated strains [8]. Finally, 86 bacterial and 22 fungal networks were reconstructed from the WGS data [9,10,11,12] (Fig. 1B).

To ensure the quality of the algal metabolic network, it is crucial to assess whether its topology enables the biosynthesis of key biomass compounds from the available environmental nutrients. This requires defining both the algal growth medium and a comprehensive biomass function that here accounts for a wide range of molecules measured in the species, including both small molecules and more complex macromolecules characteristic of its composition. In this study, since *A. nodosum* is not cultivated but rather sampled from the seashore, the composition of seawater was used as the reference growth medium. The biomass function was specifically formulated based on metabolomic data

from the literature (see below) to ensure the inclusion of a broad spectrum of known compounds. It was then mapped to the MetaCyc database [18,19] (Fig 1C).

As an initial step in functional gap-filling, the defined growth environment (seawater) was used as the input set of compounds (seeds), while the biomass function represented the target metabolites for the discrete-dynamical gap-filling tool Meneco [16] (Fig 1 GF0). This approach first identified compounds that remained non-producible even after integrating all reactions from the MetaCyc database into the *A. nodosum* GEM reconstructed with AuCoMe. In this case, 6 out of the 75 biomass target compounds were predicted to be unproducible. Their production was made possible by the addition 1 reaction from BIGG [20] intializing the L-carnitine biosynthesis pathway, 1 from KEGG [21] for the biosynthesis of linolenic acid, and 1 from Rhea [22] for the biosynthesis of 2-hydroxymyristate. For cellulose, the polymer chain extension reaction was already present, a chain initialization reaction consuming UDP-alpha-D-glucose and producing UDP and cellulose was created. Additional 76 reactions were added for the transport of the 38 seeds compounds within the cell compartments, 1 defining the biomass function and 1 for the biomass export (Fig 1D).

A second functional gap-filling step was performed to assess the number of biomass compounds that could be synthesized through reaction chains in the GEM (31 out of 75 target compounds) and to identify additional reactions from the MetaCyc v26.0 database [18,19] required to maximize the producibility of the biomass function. This process resulted in the prediction of 116 necessary reactions. In order to evaluate and filter the reactions predicted to have an added value over the production of the biomass function, several steps were performed before adding them to the network. First (Fig 1 GF1), a proteome-based gap-filling step checks for orthologs (e-value$\leq$1e-10) between UniProt protein sequences associated with MetaCyc reactions proposed and the algal proteome (via BlastP) and genome (via TBlastN). Reactions with confirmed orthologs are integrated in the network indicating the corresponding gene(s) in metadata. In our case study, 39 were included due to their potential orthologous associated proteins with *A. nodosum* proteome (identified via BlastP). Second (Fig 1 GF2), a phylogenetic gap-filling step adds candidate reactions if they are present in the GEMs of one of the 45 closely related species (brown algae); the species is indicated in the GEM metadata for the sake of tracability. Third (Fig 1 GF3), a microbiota gap-filling step adds candidate reactions found in algal microbiota GEMs and flags them as potential microbial sources. In our case-study, 20 reactions were added because they were found in brown algae and/or in the *A. nodosum* microbiota, five of them being exclusive to the microbiota. The last black-box gap-filling step (Fig 1 GF4) consists in incorporating reactions to complete the GEM without specific explanatory link to available genomes. In our case-study, this corresponded to 29 reactions. All these steps are integrated in the MeneVal pipeline which automates AuCome and Meneco executions and input files creation. The method iteratively re-executes Meneco between each validation and reaction addition step to refine the set of candidate reactions, minimizing reactions functional redundancy for limiting unnecessary reactions addition (Fig 1E).

Based on this gap-filling workflow, the initial number of 116 candidate reactions to be added to the GEM was reduced to 88, among with 49 are associated with protein sequences linked to the initial genome, and 29 reactions corresponds to biological processes for which biological evidences have to be found with specific methods.

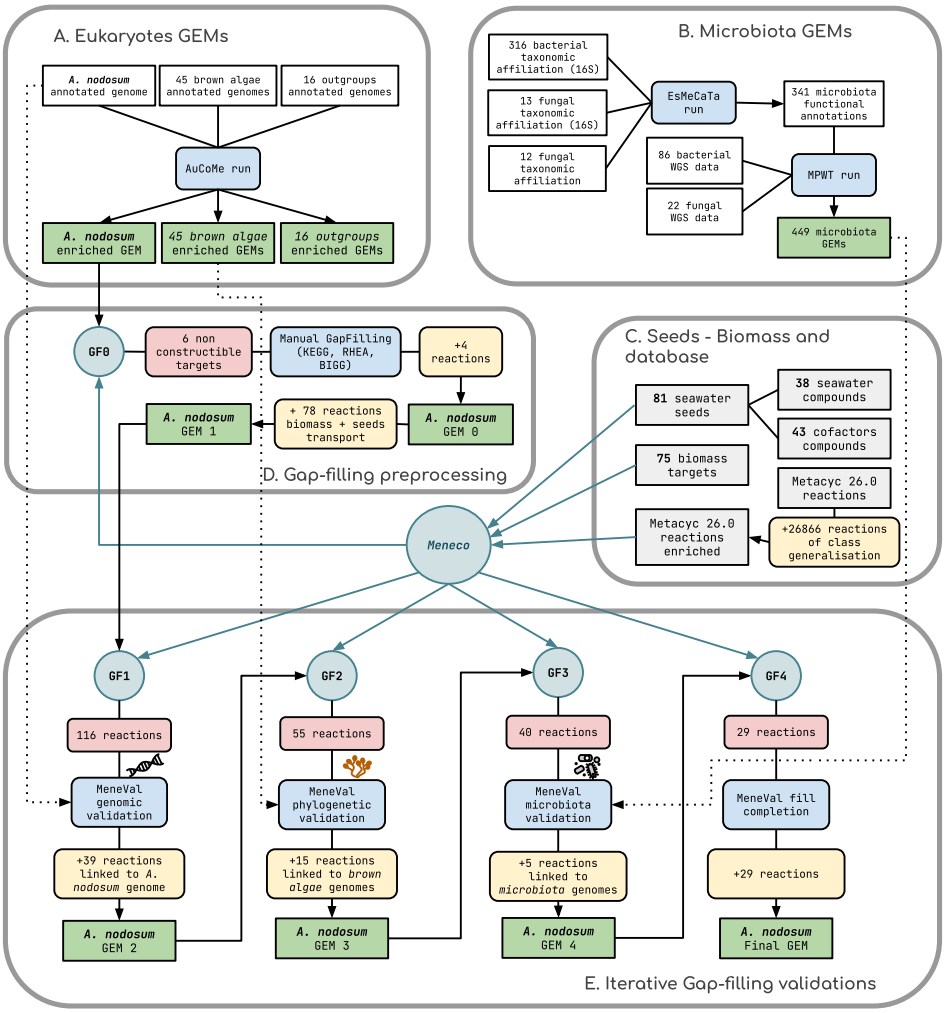


Fig. 1. *Complete pipeline for the construction and curation of our eukaryotic GEM from multiple eukaryotic genomes and associated microbiomes.*

Creation of tailor-made, species-specific algal biomass Biomass composition of *A. nodosum* was established by aggregating data from multiple sources to form a consensus. Compound measurements were primarily drawn from chemical quantification studies covering diverse conditions such as seasonality and geographic location [23,24,25] and included data from Russia's White Sea and from Bodø, Northern Norway. Additional references included core metabolism compounds from *Ectocarpus sp.7* [26] and more specific metabolites such as fucoidans [27,28,29], phlorotannins [30], and fucoxanthin [31]. Instead of focusing on specific conditions, the quantification was designed to capture general trends in the relative proportions of different compound classes, ensuring compatibility with potential future quantitative analyses. Values were estimated using averages and proportionality ratios. When specific measurements were available, averages were calculated, and quantities in µmol/g dry weight were extrapolated based on the *Ectocarpus sp.7* 20 proteinogenic amino acids ratios [26]. When only class-level data was available, the proportions of sub-compounds were estimated based on reported ratios. Regarding proteinogenic amino acids, for example, quantities were extrapolated based on the ratios observed in Ectocarpus sp. 7. When no such data were available, we assumed an equal distribution across all compounds of the class.

The resulting biomass consists of a total of 75 compounds (Fig. 2), including 2.21% of all the 20 proteinogenic amino acids, whose synthesis ensures enzymatic production autonomy. It also contains a set of 12 carbohydrates, including 6 monosaccharides (Glucose, Galactose, Xylose, Mannose, Fucose, Xylonate) and 2 more complex polysaccharides (alginate and cellulose), which make up approximately 55% of the biomass. Additionally, a set of 12 fatty acids accounts for 5% of the composition. Regarding the production of metabolites more specific to brown algae, GDP-L-fucose is included as a precursor of fucoidan [28], along with a set of monosaccharides serving as its structural components. UDP-glucose was also introduced as it a precursor in cellulose biosynthesis. For phlorotannins, only the phloroglucinol precursor [30] was incorporated, as it has a well-documented biosynthetic pathway. This final biomass function retained only major compounds, prioritizing those with known biosynthetic pathways described in databases or supported by literature.

Analysing algal biomass according to the pathways information of the predicted algal and holobiont GEMs.

Fig. 3 classifies the different compounds of the biomass according to the sources needed in the GSM reconstructon process needed to predict their production. After the initial reconstruction of the *A. nodosum* GEM using AuCoMe and before gap-filling, the model could produce 30 out of the 75 biomass compounds, primarily canonical amino acids (16 out of 20) and, more broadly, carboxylates (28 out of 30). Following gap-filling, including the addition of 4 manually curated reactions, 39 reactions from genomic validation, 15 from phylogenetic validation, and five from microbiota validation, 20 additional biomass compounds became producible, with a predominance of fatty acids, carotenoids, and more generally, carboxylates. The final set of 29 reactions added to enable the synthesis of the remaining 25 biomass compounds introduced a more diverse range of metabolites, with a higher proportion of sugars, amines, chlorophyll derivatives, and vitamins.

Discussion

Tailor-made biomass function to refine network qualitative analyses. The biomass function built in this study is based on estimated quantities derived from a consensus of multiple studies. The

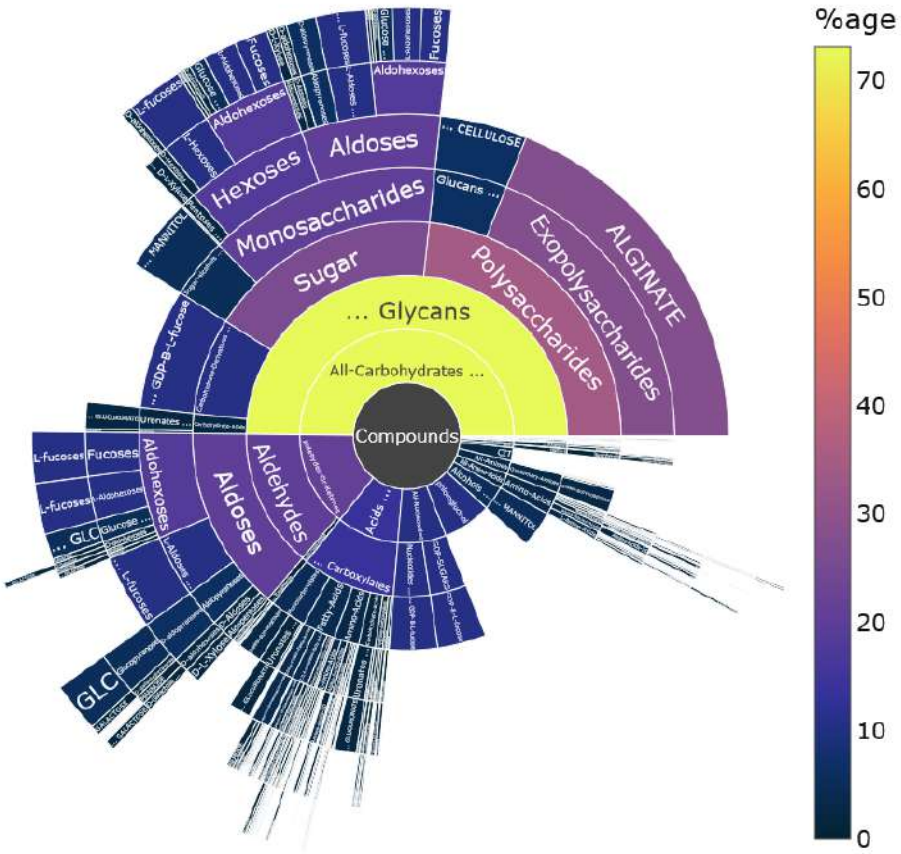


Fig. 2. *Ascophyllum nodosum biomass compounds approximated percentage quantities according to MetaCyc ontology compounds hierarchical classification.*

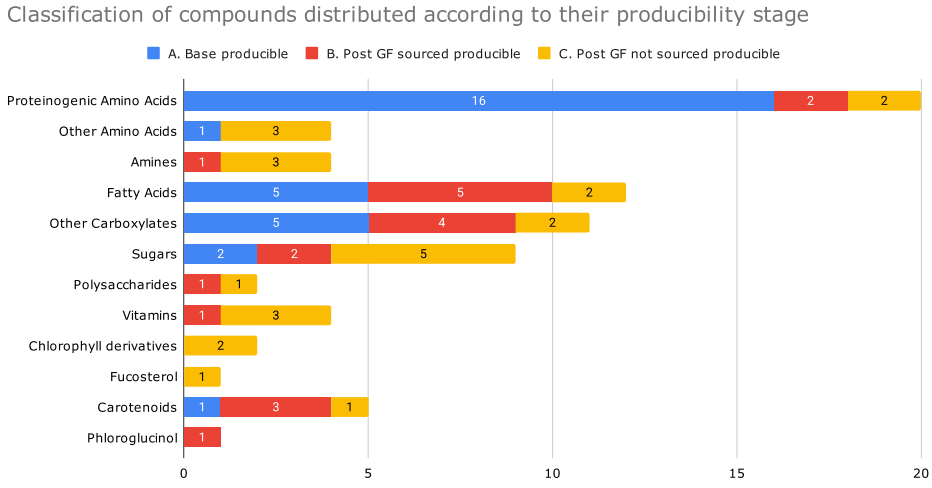


Fig. 3. *Classification of biomass compounds producible from seawater components. (GF=Gap-Filling).* ***A.*** *(blue) 30 compounds producible before gap-filling – Biosynthetic reactions linked to algal genes.* ***B.*** *(red) 20 additional compounds producible after adding 24 candidate gap-filling reactions – Supported by genomic information.* ***C.*** *(yellow) 25 remaining biomass compounds producible after adding the final 29 candidate reactions – Completing the gap-filling process.*

latter enabled the application of advanced gap-filling methods, revealing genomic information that is present or missing for the biosynthesis of a wide range of characteristic compounds. It allows for a more in-depth qualitative refinement of the GEM compared to biomass formulations previously applied in quantitative analyses applied to brown algae like *Ectocarpus sp.7* [26] and *Saccharina japonica* [32]. This approach still retains the weighted compounds necessary for such analyzes assuming the addition of RNA, DNA, and the required biosynthetic energy (ATP) [33,34]. However, these values should not be considered fixed or universal, as significant fluctuations have been reported throughout the algal life cycle, influenced by factors such as seasonality [23], temperature, and geographic location. Additionally, it is important to note that the described compounds vary in complexity, ranging from small molecules to macromolecules. As a result, the reported quantities may not always be mutually exclusive, as some compounds could be structural components of others. To take into account such variability, MDF approaches integrating ranges of metabolite concentrations [35] may be closer to the biological reality, and are indeed used for plant genome-scale modeling [36].

A Gap-filling approach to maximize the number of reactions linked to genomic information and promote explicability While the acceptable level of uncertainty in introducing reactions through gap-filling remains debated [2], the iterative reconstruction method presented here helps minimize false-positive reactions while assigning metadata to support hypotheses on their biological relevance. Notably, 39 reactions were linked to genes present in the *A. nodosum* genome but missed during automated gene prediction. An additional 15 reactions, found in other brown algal genomes, may also exist in misassembled regions of the *A. nodosum* genome. Meanwhile, five reactions originating from associated microbiota offer a limited yet valuable set of candidates for further study of potential symbiotic interactions at the holobiont level. In addition, the number of reactions added without a clear origin was reduced from 116 to 29. Despite these advantages, certain limitations remain. For validation via BlastP, not all reactions in databases are linked to protein sequences, preventing comprehensive testing. Additionally, matches within closely related enzyme families do not guarantee identical metabolic functions. Alternative approaches, such as AlphaFold, could provide deeper insights by analyzing structural-functional similarities rather than relying solely on sequence alignments.

Database knowledge dependency A limitation of GEM reconstruction based on genomic data is its reliance on existing database knowledge associating functional annotations with genomic sequences, which primarily covers well-characterized core metabolism and lacks specificity. When applied to understudied species like brown algae, these automated methods fail to capture a significant part of their unique metabolic features. This highlights a major challenge: the need to explore the "omics dark matter" [37,38]. Developing alternative methodologies capable of suggesting biosynthetic pathways for compounds without known biosynthesis pathways, particularly in poorly studied species or metabolites, will be essential to further enhance metabolic network reconstruction.

Availability and Implementation

The network reconstructed with AuCoMe before gap-filling has been made accessible via Wiki on the Phaeoexplorer project website (https://phaeoexplorer.sb-roscoff.fr/metabolic_networks/). However, the final network after gap-filling is not yet available. The code for the MeneVal pipeline is available on GitHub (https://github.com/AuReMe/MeneVal).

Acknowledgements

Our thanks go to the Phaeoexplorer project for providing the genomic data, to the members of the ANR Seabioz project for data sharing and fruitful discussions, and to the Genouest platform for providing the computational resources.

Funding information

The work was carried out as part of the Seabioz ANR program (ANR-20-CE43-0013), and has received financial support from the Centre National de la Recherche Scientifique (CNRS-MITI Algometabionte).

References

[1] Gu C, Kim GB, Kim WJ, Kim HU, Lee SY. Current Status and Applications of Genome-Scale Metabolic Models. Genome Biology. 2919;20(1):121. Available from: https://genomebiology.biomedcentral.com/articles/10.1186/s13059-019-1730-3.

[2] Bernstein DB, Sulheim S, Almaas E, Segrè D. Addressing Uncertainty in Genome-Scale Metabolic Model Reconstruction and Analysis. Genome Biology. 2021;22(1):64. Available from: https://genomebiology.biomedcentral.com/articles/10.1186/s13059-021-02289-z.

[3] Nègre D, Larhlimi A, Bertrand S. Reconciliation and Evolution of Penicillium Rubens Genome-Scale Metabolic Networks–What about Specialised Metabolism? PLOS ONE. 2023;18(8):e0289757. Available from: https://dx.plos.org/10.1371/journal.pone.0289757.

[4] Rousseau C, Demoulinger G, Rousvoal S, Champeval D, Dolly M, Michel G, et al. A Review on the Chemical Ecology of the Fucaceae Holobionts: From Fundamental Knowledge to Applications. Comptes Rendus Chimie. 2025;26(S2):23-47. Available from: https://comptes-rendus.academie-sciences.fr/chimie/articles/10.5802/crchim.271/.

[5] Belcour A, Got J, Aite M, Delage L, Collén J, Frioux C, et al. Inferring and Comparing Metabolism across Heterogeneous Sets of Annotated Genomes Using AuCoMe. Genome Research. 2023;33(6):972-87. Available from: http://genome.cshlp.org/lookup/doi/10.1101/gr.277056.122.

[6] Denoeud F, Godfroy O, Cruaud C, Heesch S, Nehr Z, Tadrent N, et al. Evolutionary Genomics of the Emergence of Brown Algae as Key Components of Coastal Ecosystems. Cell. 2024;187(24):6943-65.e39. Available from: https://linkinghub.elsevier.com/retrieve/pii/S0092867424012728.

[7] Rousseau C, Tanguy G, Legeay E, Blanquart S, Belcour A, Rousvoal S, et al. A Duo of Fungi and Complex and Dynamic Bacterial Community Networks Contribute to Shape the *AscophyllumNodosum* Holobiont; 2025. Prepublished. Available from: http://biorxiv.org/lookup/doi/10.1101/2025.03.20.643298.

[8] Vallet M, Strittmatter M, Murúa P, Lacoste S, Dupont J, Hubas C, et al. Chemically-Mediated Interactions Between Macroalgae, Their Fungal Endophytes, and Protistan Pathogens. Frontiers in Microbiology. 2018;9:3161. Available from: https://www.frontiersin.org/article/10.3389/fmicb.2018.03161/full.

[9] Bergk-Pinto B, Hamon-Giraud P, Demoulinger G, Got J, Rousvoal S, Tanguy G, et al. Exploring the Specialized Metabolism of a Brown Algal Holobiont; 2025. Unpublished. Available from: Unpublished.

[10] Martin M, Barbeyron T, Martin R, Portetelle D, Michel G, Vandenbol M. The Cultivable Surface Microbiota of the Brown Alga Ascophyllum nodosum is Enriched in Macroalgal-Polysaccharide-Degrading Bacteria. Frontiers in Microbiology. 2015;6. Available from: https://www.frontiersin.org/journals/microbiology/articles/10.3389/fmicb.2015.01487/full.

[11] Pfister C, Berlinghof J, Bogan M, Cardini U, Gobet A, Hamon-Giraud P, et al. Evolutionary history and association with seaweeds shape the genomes and metabolisms of marine bacteria; 2025. Submitted preprint.

[12] Barbeyron T, Thiébaud M, Le Duff N, Martin M, Corre E, Tanguy G, et al. Zobellia roscoffensis sp. nov. and Zobellia nedashkovskayae sp. nov., two flavobacteria from the epiphytic microbiota of the brown alga Ascophyllum nodosum, and emended description of the genus Zobellia. International Journal of Systematic and Evolutionary Microbiology. 2021;71(8):004913. Available from: https://www.microbiologyresearch.org/content/journal/ijsem/10.1099/ijsem.0.004913.

[13] Belcour A, Hamon-Giraud P, Mataigne A, Ruiz B, Le Cunff Y, Got J, et al. Estimating Consensus Proteomes and Metabolic Functions from Taxonomic Affiliations; 2022. Prepublished. Available from: http://biorxiv.org/lookup/doi/10.1101/2022.03.16.484574.

[14] Belcour A, Frioux C, Aite M, Bretaudeau A, Hildebrand F, Siegel A. Metage2Metabo, Microbiota-Scale Metabolic Complementarity for the Identification of Key Species. eLife. 2020;9:e61968. Available from: https://elifesciences.org/articles/61968.

[15] Aite M, Chevallier M, Frioux C, Trottier C, Got J, Cortés MP, et al. Traceability, Reproducibility and Wiki-Exploration for "à-La-Carte" Reconstructions of Genome-Scale Metabolic Models. PLOS Computational Biology. 2018;14(5):e1006146. Available from: https://dx.plos.org/10.1371/journal.pcbi.1006146.

[16] Prigent S, Frioux C, Dittami SM, Thiele S, Larhlimi A, Collet G, et al. Meneco, a Topology-Based Gap-Filling Tool Applicable to Degraded Genome-Wide Metabolic Networks. PLOS Computational Biology. 2017;13(1):e1005276. Available from: https://dx.plos.org/10.1371/journal.pcbi.1005276.

[17] Karp PD, Midford PE, Billington R, Kothari A, Krummenacker M, Latendresse M, et al. Pathway Tools Version 23.0 Update: Software for Pathway/Genome Informatics and Systems Biology. Briefings in Bioinformatics. 2021;22(1):109-26. Available from: https://academic.oup.com/bib/article/22/1/109/5669859.

[18] Caspi R, Billington R, Keseler IM, Kothari A, Krummenacker M, Midford PE, et al. The MetaCyc Database of Metabolic Pathways and Enzymes - a 2019 Update. Nucleic Acids Research. 2020;48(D1):D445-53. Available from: https://academic.oup.com/nar/article/48/D1/D445/5581728.

[19] Caspi R, Dreher K, Karp PD. The Challenge of Constructing, Classifying, and Representing Metabolic Pathways. FEMS Microbiology Letters. 2013;345(2):85-93. Available from: https://academic.oup.com/femsle/article-lookup/doi/10.1111/1574-6968.12194.

[20] King ZA, Lu J, Dräger A, Miller P, Federowicz S, Lerman JA, et al. BiGG Models: A Platform for Integrating, Standardizing and Sharing Genome-Scale Models. Nucleic Acids Research. 2016;44(D1):D515-22. Available from: https://academic.oup.com/nar/article-lookup/doi/10.1093/nar/gkv1049.

[21] Kanehisa M. KEGG: Kyoto Encyclopedia of Genes and Genomes. Nucleic Acids Research. 2000;28(1):27-30. Available from: https://academic.oup.com/nar/article-lookup/doi/10.1093/nar/28.1.27.

[22] Bansal P, Morgat A, Axelsen KB, Muthukrishnan V, Coudert E, Aimo L, et al. Rhea, the Reaction Knowledgebase in 2022. Nucleic Acids Research. 2022;50(D1):D693-700. Available from: https://academic.oup.com/nar/article/50/D1/D693/6424769.

[23] Bogolitsyn K, Parshina A, Ivanchenko N, Polomarchuk D. Seasonal Variations in the Chemical Composition of Arctic Brown Macroalgae. Algal Research. 2023;72:103112. Available from: https://linkinghub.elsevier.com/retrieve/pii/S2211926423001455.

[24] Belghit I, Rasinger JD, Heesch S, Biancarosa I, Liland N, Torstensen B, et al. In-Depth Metabolic Profiling of Marine Macroalgae Confirms Strong Biochemical Differences between Brown, Red and Green Algae. Algal Research. 2017;26:240-9. Available from: https://linkinghub.elsevier.com/retrieve/pii/S2211926417303880.

[25] Demoulinger G, Rousseau C, Leblanc C, Michel G, Thomas F, Markov GV, et al. Metabolites of Ascophyllum nodosum: Unlocking Undustrial Potential; 2025. Unpublished.

[26] Prigent S, Collet G, Dittami SM, Delage L, Ethis De Corny F, Dameron O, et al. The Genome-scale Metabolic Network of *Ectocarpus Siliculosus* (Ecto GEM): A Resource to Study Brown Algal Physiology and Beyond. The Plant Journal. 2014;80(2):367-81. Available from: https://onlinelibrary.wiley.com/doi/10.1111/tpj.12627.

[27] Li B, Lu F, Wei X, Zhao R. Fucoidan: Structure and Bioactivity. Molecules. 2008;13(8):1671-95. Available from: https://www.mdpi.com/1420-3049/13/8/1671.

[28] Skriptsova AV. Fucoidans of Brown Algae: Biosynthesis, Localization, and Physiological Role in Thallus. Russian Journal of Marine Biology. 2015;41(3):145-56. Available from: http://link.springer.com/10.1134/S1063074015030098.

[29] Fu C, Xu X, Xie Y, Liu Y, Liu M, Chen A, et al. Rational Design of GDP-d-mannose Mannosyl Hydrolase for Microbial L-fucose Production. Microbial Cell Factories. 2023;22(1):56. Available from: https://microbialcellfactories.biomedcentral.com/articles/10.1186/s12934-023-02060-y.

[30] Meslet-Cladière L, Delage L, Leroux CJJ, Goulitquer S, Leblanc C, Creis E, et al. Structure/Function Analysis of a Type III Polyketide Synthase in the Brown Alga *Ectocarpus Siliculosus* Reveals a Biochemical Pathway in Phlorotannin Monomer Biosynthesis. The Plant Cell. 2013;25(8):3089-103. Available from: https://academic.oup.com/plcell/article/25/8/3089-3103/6096485.

[31] Cunningham EM, O'Kane AP, Ford L, Sheldrake GN, Cuthbert RN, Dick JTA, et al. Temporal Patterns of Fucoxanthin in Four Species of European Marine Brown Macroalgae. Scientific Reports. 2023;13(1):22241. Available from: https://www.nature.com/articles/s41598-023-47274-7.

[32] Nègre D, Aite M, Belcour A, Frioux C, Brillet-Guéguen L, Liu X, et al. Genome–Scale Metabolic Networks Shed Light on the Carotenoid Biosynthesis Pathway in the Brown Algae Saccharina Japonica and Cladosiphon Okamuranus. Antioxidants. 2019;8(11):564. Available from: https://www.mdpi.com/2076-3921/8/11/564.

[33] Feist AM, Palsson BO. The biomass objective function. Current Opinion in Microbiology. 2010;13(3):344-9. Available from: https://www.sciencedirect.com/science/article/pii/S1369527410000512.

[34] Kim J, Fabris M, Baart G, Kim MK, Goossens A, Vyverman W, et al. Flux balance analysis of primary metabolism in the diatom *Phaeodactylum tricornutum*. The Plant Journal. 2016;85(1):161-76. Available from: https://onlinelibrary.wiley.com/doi/10.1111/tpj.13081.

[35] Noor E, Bar-Even A, Flamholz A, Reznik E, Liebermeister W, Milo R. Pathway thermodynamics highlights kinetic obstacles in central metabolism. PLoS computational biology. 2014;10(2):e1003483.

[36] Chowdhury NB, Simons-Senftle M, Decouard B, Quillere I, Rigault M, Sajeevan KA, et al. A multi-organ maize metabolic model connects temperature stress with energy production and reducing power generation. iScience. 2023;26(12):108400. Available from: https://linkinghub.elsevier.com/retrieve/pii/S258900422302477X.

[37] Monge ME, Dodds JN, Baker ES, Edison AS, Fernández FM. Challenges in Identifying the Dark Molecules of Life. Annual Review of Analytical Chemistry. 2019;12(1):177-99. Available from: https://www.annualreviews.org/doi/10.1146/annurev-anchem-061318-114959.

[38] Ellens KW, Christian N, Singh C, Satagopam VP, May P, Linster CL. Confronting the catalytic dark matter encoded by sequenced genomes. Nucleic Acids Research. 2017;45(20):11495-514. Available from: http://academic.oup.com/nar/article/45/20/11495/4559119.

Predictive modelling of Acute Promyelocytic Leukaemia resistance to Retinoic Acid therapy.

José A. SANCHEZ-VILLANUEVA[1], Lia N'GUYEN[2], Mathilde POPLINEAU[2], Estelle DUPREZ[2,#], Élisabeth REMY[1,#], Denis THIEFFRY[3,4,#]

1 Institute of Mathematics of Marseilles, Aix Marseille University, CNRS, 163 Av.de Luminy, 13009 Marseille, France.

2 Integrative molecular biology in hematopoiesis and leukemia, Equipe Labellisée Ligue Contre le Cancer, CRCM, Inserm UMR1068, CNRS UMR7258, Institut Paoli-Calmettes, Aix Marseille University, 232 Bd de Sainte-Marguerite, 13009 Marseille, France.

3 Department of Biology, École Normale Supérieure, PSL Research University, 46 rue d'Ulm, 75005 Paris, France.

4 Institut Curie - INSERM U900 - Mines Paris, PSL Research University, 11-13 Rue Pierre et Marie Curie, 75005 Paris, France.

Corresponding Authors: estelle.duprez@inserm.fr, elisabeth.remy@univ-amu.fr, denis.thieffry@ens.fr

Paper Reference: Sánchez-Villanueva *et al.* (2025) Dynamical modelling of the regulatory network underlying Retinoic Acid resistance in Acute Promyelocytic Leukaemia. *Briefings in Bioinformatics* 26: bbaf002. https://doi.org/10.1093/bib/bbaf002

Keywords

Regulatory network, Logical model, Acute Promyelocytic Leukaemia, Epigenetic regulation, Therapy resistance.

Abstract

Acute Promyelocytic Leukaemia (APL) arises from an aberrant chromosomal translocation involving the Retinoic Acid Receptor Alpha (RARA) gene, predominantly with the Promyelocytic Leukaemia (PML) or Promyelocytic Leukaemia Zinc Finger (PLZF) genes. The resulting oncoproteins block the haematopoietic differentiation program promoting aberrant proliferative promyelocytes. Retinoic Acid (RA) therapy is successful in most of the PML::RARA patients, while PLZF::RARA patients frequently become resistant and relapse.

Recent studies pointed to various underlying molecular components, but their precise contributions remain to be deciphered. We developed a logical network model integrating signalling, transcriptional and epigenetic regulatory mechanisms, which captures key features of the APL cell responses to RA depending on the genetic background.

The explicit inclusion of the histone methyltransferase EZH2 allowed the assessment of its role in the resistance mechanism, distinguishing between its canonical and non-canonical activities.

The model dynamics was thoroughly analysed using tools integrated in the public software suite maintained by the CoLoMoTo consortium (https://colomoto.github.io/). The model serves as a solid basis to assess the roles of novel regulatory mechanisms, as well as to explore novel therapeutical approaches in silico.

Highlight

This article presents a predictive logical model capturing and explaining key features of the regulatory network underlying the responses of two main subtypes of Acute Promyelocytic Leukaemia (APL) cells to Retinoic Acid (RA) therapy.

The stable states of the model recapitulate the phenotypes of differentiated and aberrant proliferative cells induced by RA treatment for two different APL genetic backgrounds, while a commitment analysis identifies the crucial components underlying the decision between cell differentiation and aberrant proliferation.

The simulations of different EZH2 perturbations and a parameter sensitivity analysis enables the characterisation of the components of the network underlying cell fate decisions, distinguishing the canonical versus non-canonical activities of EZH2, highlighting the key role of the non-canonical activity of EZH2 in the maintenance of the resistance to RA treatment, and pointing to potential targets for novel combinatorial therapy strategies.

Finally, the model analysis relies on a robust computational workflow combining four different tools, developed by different groups but seamlessly integrated in the CoLoMoTo software environment (https://colomoto.github.io/). The integration of these tools in a common framework together with the use of Jupyter notebooks foster the reproducibility of our computational results, and simultaneously ease further refinements or extensions of this modelling study. The software GINsim, the CoLoMoTo environment, as well as the model file and the Jupyter notebook enabling the reproduction of the results of this study are all available online (http://ginsim.org/node/256).

Building a modular and multi-cellular virtual twin of the synovial joint in Rheumatoid Arthritis

Naouel ZERROUK[1,2*], Franck AUGE[1*] & Anna NIARAKIS[1,3*]

1 Sanofi R&D Data and Data Science, Artificial Intelligence & Deep Analytics, Omics Data Science, Chilly-Mazarin, France

2 GenHotel, Laboratoire Européen de Recherche Pour La Polyarthrite Rhumatoïde, University Paris-Saclay, University Evry, Evry, France

3 Lifeware Group, Inria Saclay, Palaiseau, France

Corresponding Author: anna.niaraki@univ-tlse3.fr

* current address for Anna Niarakis: University of Toulouse, Department of Biology and Geosciences, Faculty of Sciences and Engineering, MCD-CBI, Toulouse

* current address for Franck Augé: Servier Paris Saclay, R&D

* current address for Naouel Zerrouk: Cure51, 19 Rue Richer, 75009 Paris

Paper Reference: Zerrouk, N., Augé, F. & Niarakis, A. Building a modular and multi-cellular virtual twin of the synovial joint in Rheumatoid Arthritis. npj Digit. Med. 7, 379 (2024).
https://doi.org/10.1038/s41746-024-01396-y

Keywords

Rheumatoid arthritis, large scale Boolean model, virtual twin

Abstract

Rheumatoid arthritis is a complex disease marked by joint pain, stiffness, swelling, and chronic synovitis, arising from the dysregulated interaction between synoviocytes and immune cells. Its unclear etiology makes finding a cure challenging. The concept of digital twins, used in engineering, can be applied to healthcare to improve diagnosis and treatment for complex diseases like rheumatoid arthritis. In this work, we pave the path towards a digital twin of the arthritic joint by building a large, modular biochemical reaction map of intra- and intercellular interactions. This network, featuring over 1000 biomolecules, is then converted to one of the largest executable Boolean models for biological systems to date. Validated through existing knowledge and gene expression data, our model is used to explore current treatments and identify new therapeutic targets for rheumatoid arthritis.

Highlight

Digital twin implementation in healthcare has the potential to advance biomedical research with applications for personalised medicine, pharmaceutical development, and clinical trials [1]. Current tangible implementations of digital twins can be found in precision cardiology [2], type 1 diabetes [3], cancer [4], and epidemic outbreaks [5]. In these applications, researchers combine several cutting-edge technologies, including mathematical modelling. This work initiates the development of a virtual twin of the arthritic joint, first by constructing a comprehensive large-scale map that depicts both the intra- and intercellular interactions involved in RA pathogenesis. The map incorporates the four cell-specific maps of the RA Atlas [6], describing the synovial fibroblast, M1 and M2 macrophages, and CD4 + Th1 cell types.

Furthermore, it integrates bidirectional cellular communication between these cell types, providing a detailed multicellular representation of the RA synovium. The map is modular, allowing for future expansion with additional cell-specific maps. We employed the Boolean formalism to explore the system's emergent behaviour. Boolean models can handle large-scale systems and do not require quantitative parameters. We used the map to model translation framework and the tool CaSQ described in Aghamiri et al., 2020 [7] to translate the multicellular map to a fully executable, large-scale Boolean model. The dynamic behaviour of the RA multi-cellular model was tested against prior knowledge to assess its capacity to reproduce known biological mechanisms. The RA multi-cellular model is significantly larger in scale compared to the two macrophage models tested in Zerrouk et al., 2024 [8], demonstrating the scalability of the proposed computational framework. The model was then used to study the mechanism of action of current RA treatments and identify new potential therapeutic targets and drug combinations via single- and double-knockout *in silico* simulations.

References

1. National Academies of Sciences, Engineering, and Medicine; National Academy of Engineering; Division on Earth and Life Studies; Division on Engineering and Physical Sciences; Board on Life Sciences; Board on Atmospheric Sciences and Climate; Computer Science and Telecommunications Board; Board on Mathematical Sciences and Analytics. Opportunities and challenges for digital twins in atmospheric and climate sciences: proceedings of a workshop—in brief. Casola L, editor. Washington (DC): National Academies Press (US); 2023.
2. Corral-Acero J, Margara F, Marciniak M, Rodero C, Loncaric F, Feng Y, et al. The “Digital Twin” to enable the vision of precision cardiology. Eur Heart J. 2020 Dec 21;41(48):4556–64.
3. Breton MD, Kanapka LG, Beck RW, Ekhlaspour L, Forlenza GP, Cengiz E, et al. A Randomized Trial of Closed-Loop

Control in Children with Type 1 Diabetes. N Engl J Med. 2020 Aug 27;383(9):836–45.

4. Batch KE, Yue J, Darcovich A, Lupton K, Liu CC, Woodlock DP, et al. Developing a cancer digital twin: supervised metastases detection from consecutive structured radiology reports. Front Artif Intell. 2022 Mar 2;5:826402.

5. Ivanov D. Predicting the impacts of epidemic outbreaks on global supply chains: A simulation-based analysis on the coronavirus outbreak (COVID-19/SARS-CoV-2) case. Transp Res E Logist Transp Rev. 2020 Apr;136:101922.

6. Zerrouk N, Aghakhani S, Singh V, Augé F, Niarakis A. A mechanistic cellular atlas of the rheumatic joint. Front Syst Biol. 2022 Jul 11;2.

7. Aghamiri SS, Singh V, Naldi A, Helikar T, Soliman S, Niarakis A. Automated inference of Boolean models from molecular interaction maps using CaSQ. Bioinformatics. 2020 Aug 15;36(16):4473–82.

8. Zerrouk N, Alcraft R, Hall BA, Augé F, Niarakis A. Large-scale computational modelling of the M1 and M2 synovial macrophages in rheumatoid arthritis. NPJ Syst Biol Appl. 2024 Jan 26;10(1):10.

Session 4: Structural Bioinformatics and Proteomics

Searching for variable structural motifs in RNA graphs using simple descriptors

Camille DE AMORIM[1] and Alain DENISE[1,2]

1 Université Paris-Saclay, CNRS, Laboratoire Interdisciplinaire des Sciences du Numérique (LISN), Orsay, 91405, France

2 Université Paris-Saclay, CEA, CNRS, Institute for Integrative Biology of the Cell (I2BC), Orsay, 91198, France

Corresponding author: alain.denise/camille.de-amorim@universite-paris-saclay.fr

Abstract

RNA tertiary structure prediction is a problem that is far from being solved, as shown in [1]. The tertiary structure depends very strongly on particular types of interactions, known as non-canonical interactions. These interactions form dense networks, called structural motifs, which recur in RNA structures and are characteristic of certain local three-dimensional shapes [2]. There are several families of such motifs; some of them are well known, such as the kink-turn and the A-minor families. RNA structures can be represented in a strictly topological manner, i.e., without taking geometric information into account, by labeled graphs in which the vertices are the nucleotides and the edges or arcs are the interactions between them, labeled by the kind of interaction. In recent years, several studies have focused on discovering, classifying, and predicting structural motifs in RNA molecules represented by such graphs. In this way, a close relationship can be established between the topological representation and the tertiary structure of RNA, taking a step towards predicting the latter. One of the main difficulties of such studies is the large variability of motifs within the same family: subgraphs corresponding to members of the same family can be rather different although their 3D shapes are similar. And, in fact, only a few works attempted to tackle this variability problem by developing sophisticated fuzzy subgraph search algorithms in RNA graphs [3,4].

References

[1] Kwon D. RNA function follows form – why is it so hard to predict? Nature. 2025 Mar;639:1106-8. Available from: https://www.nature.com/articles/d41586-025-00920-8.

[2] Leontis NB, Lescoute A, Westhof E. The building blocks and motifs of RNA architecture. Current Opinion in Structural Biology. 2006 Jun;16(3):279-87.

[3] Oliver C, Mallet V, Philippopoulos P, Hamilton WL, WaldispÅNuhl J. Vernal: a tool for mining fuzzy network motifs in RNA. Bioinformatics. 2022;38(4):970-6. Publisher: Oxford University Press. Available from: https://academic.oup.com/bioinformatics/article-abstract/38/4/970/6428528.

[4] Boury T, Ponty Y, Reinharz V. Automatic exploration of the natural variability of RNA noncanonical geometric patterns with a parameterized sampling technique. In: WABI 2023 - 23rd Workshop on Algorithms in Bioinformatics. Houston, United States: Texas A&M University; 2023. Available from: https://hal.science/hal-04094288.

Comparative Analysis of Deep Learning-Based Algorithms for Peptide Structure Prediction

Clément SAUVESTRE[1,2], Jean-François ZAGURY[2] and Florent LANGENFELD[1,2]

[1] Laboratoire GBCM, EA7528, Conservatoire national des arts et métiers (CNAM), HESAM Université, 2 Rue Conté, 75003 Paris, France

[2] Peptinov, 29 Rue Du Faubourg Saint-Jacques, 75014 Paris, France

Corresponding author: clement.sauvestre.1@gmail.com

Keywords peptide, deep learning, three-dimensional structure, prediction

Abstract *While of primary importance in both the biomedical and therapeutic fields, peptides suffer from a relative lack of dedicated tools to predict efficiently and accurately their 3D structures despite a crucial step in understanding their physio-pathological function or designing new drugs. In recent years, deep-learning methods have enabled a major breakthrough for the protein 3D structures prediction approaches, allowing to predict protein 3D structures with a near-experimental accuracy for nearly any protein sequence. This present study aims at confronting some of these new methods (AlphaFold2, RoseTTAFold2 and ESMFold) for the peptides 3D structure prediction problem, and to evaluate their performance. All methods produced high quality results, but their overall performance is lower as compared to the prediction of proteins 3D structure. We also identified a few structural features that impede the ability to produce high-quality peptide structure predictions. These findings point out the discrepancy that still exists between the protein and peptide 3D structure prediction methods, and underline a few cases where the generated peptide structures should be used very cautiously.*

Introduction

Chemically, peptides are short polymers of amino acids (typically ranging from 2 to 50 residues), and their structure may exhibit diverse structural features: they can be highly flexible (*i.e.* have multiple low-energy conformations) with transient meta-stable secondary structures, display multiple cycles, *etc*. Characterizing the three-dimensional structure of peptides is essential to understanding their functions or predicting their biological effects. This structural knowledge is also important for designing effective and specific therapeutic peptides capable of modulating biological processes [1].

Computational methods have emerged as complementary tools for studying the three-dimensional structure of biological macromolecules, such as proteins and nucleic acids. Recent deep learning algorithms have demonstrated their ability to accurately predict the three-dimensional structure of proteins [2]. Among these, AlphaFold2 [3,4] has already been compared to alternative computational methods dedicated to peptide structure prediction and has outperformed them [5].

In this article, our objective is to extend previous work by comparing the performance of several non-peptide-specific algorithms to predict the three-dimensional structure of peptides. More specifically, we have chosen to evaluate the performance of AlphaFold2 (AF2) [3,4], RoseTTAFold2 (RF2) [6], and ESMFold (ESMF) [7,8], three of the most popular open-source algorithms. The purpose is not only to compare the performance of each algorithm, but also to better decipher the features associated with lower/higher quality predictions, and their current limitations. To achieve this goal, we selected

a dataset of 765 peptides from the Protein Data Bank [9,10] (PDB) and performed structure prediction calculations with each three-dimensional structure prediction algorithm. Then, we analyzed the quality of predicted structures considering various structural features such as the number of cyclizations, the type of secondary structure, the size, *etc*. To carry out this work, we used open-source tools that not only enable large-scale structure prediction but also provide for advanced users the flexibility to modify algorithms. In this work, our goal is to identify the current limitations of these methods in terms of structural features as well as to highlight their successes despite being initially designed for proteins. These results may also help the development of new methods specifically dedicated to peptide structure prediction, by leveraging the limitations observed in this work.

Methods

Creation of the peptide dataset

All chains of the Protein Data Bank [9,10] (PDB as of October 24, 2023) were clustered using MMSeqs2 [11] using settings adapted to small sequences and a similarity threshold of 70%, resulting in 67,623 clusters of 3D structures with distinct sequences. We retained monomeric protein structures from 10 to 50 residues solved by solution NMR spectroscopy, excluding non-standard residues (*e.g.* post-translational modifications, non-canonical residues, D-amino acids, *etc*). Peptides belonging to PDBTM [12], mpstruc [13], MemProtMD [14] and OPM [15] membrane databases were excluded.

As a result, a total of 765 structures were included in our dataset. This selection process ensures that only a minimal number of peptides are likely to be present in the training datasets of the compared methods [3,4,6,7,8].

Three-dimensional peptide structure prediction algorithms

AlphaFold2 (AF2), RoseTTAFold2 (RF2), and ESMFold (ESMF) – three of the most popular and effective tools for protein structure prediction at the time our study began – were evaluated (specific versions are shown in Tab. 1). AF2 and RF2 use evolutionary patterns extracted from Multiple Sequence Alignments (MSA) to derive spatial relationships between residues. These alignment-based algorithms deeply integrate the MSA into the neural network architecture through an attention mechanism to iteratively refine the predicted structure. The databases used by AF2 and RF2 to construct MSA and identify structural templates are presented in Tab. 2. ESMF uses a language model that enables rapid and accurate prediction directly from a unique sequence, capturing evolutionary couplings without using MSA while simplifying neural architecture and reducing alignment-based pipeline costs. For each peptide predicted by AF2, we selected the model with the highest average plDDT, the confidence score used by all the methods compared. This choice allows a fair comparison, as RF2 and ESMF produce a single model by default.

Tool	Version	Download date	Link
AlphaFold2	v2.3	05 Apr 2023	https://github.com/google-deepmind/alphafold
RoseTTAFold2	v1.0	12 Apr 2024	https://github.com/uw-ipd/RoseTTAFold2
ESMFold	v1.0	Nov 2022	https://github.com/facebookresearch/esm

Tab. 1. Summary of structure prediction algorithms compared.

Databases	Version — Download date	Used by
BFD	Only version available	AF2 & RF2
MGnify	v2022_05	AF2
UniRef30	v2021_03	AF2 & RF2
UniRef90	v2022_05	AF2
PDB	2023-02-09	AF2
PDB70	v2020Apr01	AF2
PDB100	v2021Mar03	RF2

Tab. 2. Sequence and structural databases used by AF2 and RF2.

Structural-based features analysis

The GDT_TS score [16] was used to assess the similarity between predicted peptide models and experimental structures obtained by NMR. Each predicted model was superposed to each reference NMR model and scored using the Local Global Alignment (LGA) algorithm [16]. The highest GDT_TS was retained for each predicted model.

We examined the quality of predictions (GDT_TS scores) according to the (a) number of disulfide bridges, (b) sequence length, (c) secondary structure topology and (d) flexibility as estimated by the average lDDT-Cα computed between all models of the NMR reference structures ($lDDT_{NMR}$).

Statistical comparisons of GDT_TS scores between peptide groups (*e.g.* 0 *versus* 3 disulfide bridges) were performed using Kruskal-Wallis H tests (one-way ANOVA on ranks) followed by Dunn's post-hoc tests to assess significant differences between groups. Wilcoxon signed-rank tests were used to compare scores between algorithms within each peptide class (*e.g.*, GDT_TS scores for peptides with 30 to 39 residues predicted by AF2 and RF2).

In our dataset, 436 peptides are linear, while 329 peptides are cyclized by disulfide bridges. For linear peptides, we evaluated whether disulfide bridges were found in the predicted models. For peptides with disulfide bridges, we evaluated whether the predictions reproduced these bridges as observed in the experimental NMR structures. We applied a threshold of 3 Å between the SG atoms of cysteine residues to determine whether the prediction was compatible with a disulfide bridge.

For AF2 and RF2, previous studies have shown that the quality of the MSA can significantly influence the accuracy of the predicted structures. To explore this effect, we analyzed the relationship between MSA quality, evaluated by the number of effective sequences (Neff) [17], and the final model accuracy, measured using the GDT_TS.

Results and Discussion

Predictive methods produce good-quality models consistently, on average.

AF2 and RF2 displayed comparable overall results, and achieved higher average, median and minimum GDT_TS values than ESMF (Tab. 3): p $= 1.00 \cdot 10^{-31}$ for AF2 *versus* ESMF; p $= 2.70 \cdot 10^{-18}$ for RF2 *versus* ESMF. AF2 and RF2 showed a small but significant difference in GDT_TS values (p $= 0.033$). Spearman's correlation tests indicate a strong correlation between GDT_TS scores for AF2, RF2, and ESMF predictions: $\rho = 0.822$ between the GDT_TS scores of AF2 and RF2, $\rho = 0.831$ between AF2

and ESMF and $\rho = 0.74$ between RF2 and ESMF. Overall, Spearman's correlations indicate that all algorithms are able to generate predictions that are consistent with each other.

Summary statistics	AF2	RF2	ESMF
Average	75.63	75.73	71.88
Median	79.17	78.57	74.31
Standard Deviation	17.11	16.43	17.89
Minimum	26.14	27.27	26.09
Maximum	100.00	100.00	100.00

Tab. 3. Descriptive statistics of the GDT_TS scores for the models predicted by AF2, RF2 and ESMF over our dataset of 765 peptides.

Our results show that AF2 and RF2, two methods based on MSA, outperform ESMF, an approach relying solely on polypeptide sequences. These results align with those obtained for the prediction of protein structures, where AF2 and RF2 usually perform better than ESMF [3,6,8]. However, while these MSA-based algorithms excel in predicting monomeric proteins (median GDT_TS $>$ 90) [3,6], their performance decreases for the peptides from our dataset (median GDT_TS $\simeq$ 79, Tab. 3). All three methods were mostly trained on sequences longer than peptide sequences (*i.e.*, $>$50 residues). This may explain the lower performance of these algorithms in predicting peptide structures compared to protein structures.

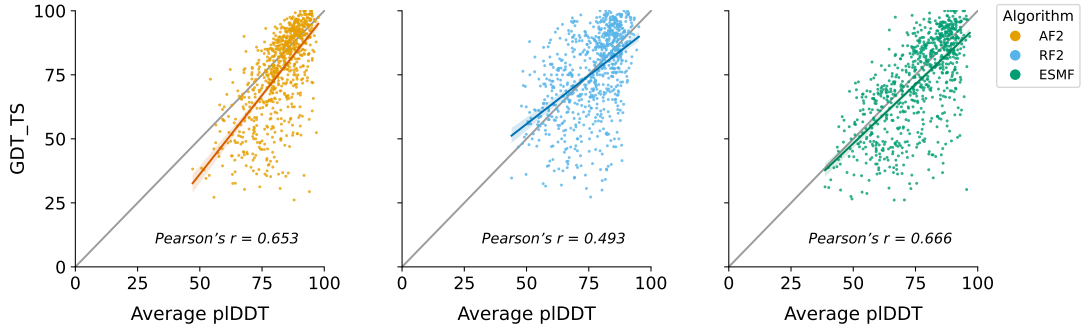


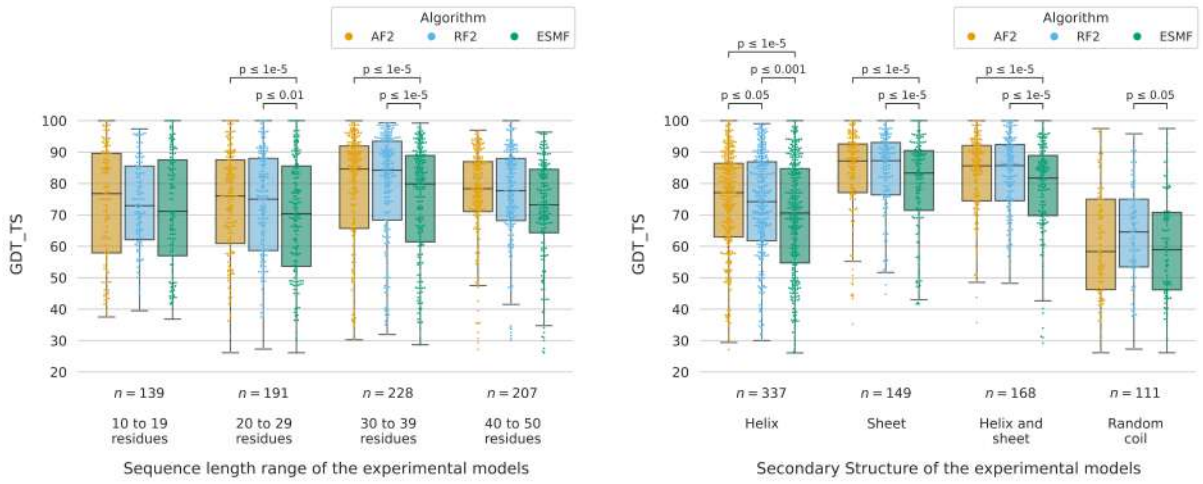
Fig. 1. Comparison of the average plDDT and GDT_TS scores of the dataset models predicted by AF2 (left panel), RF2 (middle panel) and ESMF (right panel).

We found a moderate, positive correlation between GDT_TS and average plDDT for all methods (Fig. 1), suggesting that the plDDT score is overall reliable, and can be used to estimate the accuracy of the peptide structure predictions.

Detailed analysis based on peptide structural features.

Peptide length. GDT_TS scores improve with increasing peptide size for AF2 and RF2, with the higher scores observed for peptides in the class of size 30–39 residues. In contrast, ESMF predictions are hardly affected by peptide sequence length. AF2 and RF2 consistently produced higher GDT_TS scores than ESMF for peptides in the 20–29 and 30–39 residues ranges (Fig. 2a).

Peptide secondary structure topology. Random coil (RC) peptides achieved significantly lower GDT_TS scores than peptides with a defined secondary structure (helix – H, sheet – S or mixed helix and sheet – H+S). Among structured peptides, helix-bearing peptides scored lower than either



(a) Comparison of GDT_TS scores based on peptide sequence length.

(b) Comparison of GDT_TS scores based on peptide secondary structures.

Fig. 2. Comparison of GDT_TS scores based on peptide sequence length ranges (a) and peptide secondary structures (b). Wilcoxon signed-rank tests p-values below 0.05 are indicated on the top.

sheet peptides or peptides with helices and sheets. RF2 consistently outperformed ESMF in terms of GDT_TS for all types of secondary structure (Fig. 2b). AF2 also showed significantly higher GDT_TS values than ESMF for peptides with helices, sheets or combined H+S secondary structures. Additionally, AF2 performed better than RF2 for helical peptides (Fig. 2b). This discrepancy in performance for portions of peptides structured in helix and sheet is in line with the results observed for proteins during CASP14 [18].

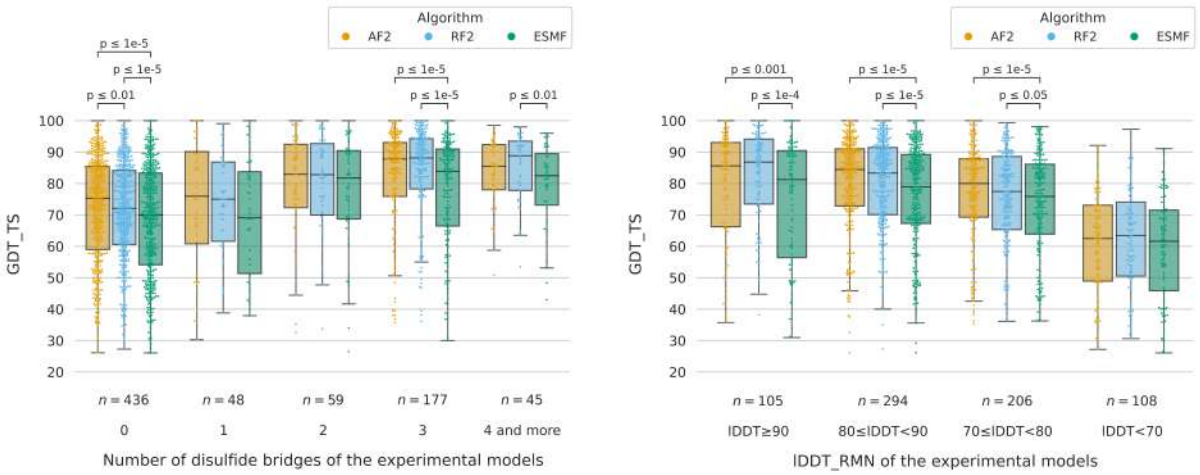
Disulfide bridge-based cyclization. We observed that AF2's predictions are more consistent with the experimentally observed disulfide bridges, followed by RF2's predictions, and finally ESMF's (Tab. 4). All methods found putative disulfide bridges in the 436 linear peptides (21, 20 and 24 for AF2, RF2 and ESMF, respectively).

Predicted models for peptides with two or more disulfide bridges exhibited significantly higher GDT_TS scores than linear peptides. For AF2 and RF2, peptides cyclized by three or more disulfide bridges also displayed higher scores than those cyclized by a single bridge. For peptides with 0, 3 and 4+ disulfide bridges, RF2 outperformed ESMF in terms of GDT_TS scores (Fig. 3a). Similarly, AF2 demonstrated greater predictive accuracy than ESMF for peptides with 0 and 3 disulfide bridges. AF2 and

Disulfide bridge	Number of peptides	AF2	RF2	ESMF
1	48	35	32	30
2	59	44	38	39
3	177	151	136	124
$\geq$4	45	39	40	35
Total	**329**	**269**	**246**	**228**

Tab. 4. Comparison of the number of experimental disulfide bridges found in structures predicted by three different algorithms. Only predicted models consistent with all experimental disulfide bridges were considered as successful.

RF2 showed comparable performance in most categories, with the exception of linear peptides, for which RF2's GDT_TS scores were significantly lower than those of AF2 (Fig. 3a).



(a) Comparison of GDT_TS scores based on disulfide bridges number.

(b) Comparison of GDT_TS scores based on peptide flexibility.

Fig. 3. Comparison of GDT_TS scores based on peptide disulfide bridges number (a) and peptide flexibility (b). Wilcoxon signed-rank tests p-values below 0.05 are indicated on the top.

Peptide flexibility. Flexible peptides ($lDDT_{NMR} < 70$) displayed significantly lower GDT_TS scores. Among peptides with $lDDT_{NMR}$ values between 70 and 100, AF2 and RF2 outperformed ESMF in terms of GDT_TS scores (Fig. 3b). These results are in line with the observations made during CASP14 [2], where flexible regions of proteins were predicted with lower quality.

Advances have already been made to overcome these limitations. Several methods [19,20] can predict alternative conformations of proteins *via* MSA clustering, for instance. Such new methods could benefit peptide structure predictions as well.

Peptides with structural features similar to proteins' are associated with high-quality predictions. Peptides with lesser flexibility are often larger (in terms of the number of amino acids), more structured (with a greater number of secondary structure elements) and cyclized through multiple disulfide bridges. All these features are associated with higher quality predictions (Fig. 2, 3) and resemble the features found in the protein structures used to train AF2, RF2, and ESMF algorithms: the more protein-like the peptide, the better the predicted model. This overall observation was expected given the training datasets used by these methods. These results underline the difficulty these methods have in extending their field of application beyond structures similar to those in their training data.

The quality of the MSA drives the quality of the predicted model.

Peptides 6CKF and 7YRW (Fig. 4) share similar structural features (mainly helical, 30-39 residues, three disulfide bridges, $lDDT_{NMR}$ 80-90), but their prediction accuracy differs. 6CKF is well-predicted (GDT_TS > 95), while 7YRW is poorly predicted (GDT_TS < 50) by all methods. This discrepancy may be linked to the quality of the MSA: the 6CKF sequence resulted in a deep alignment (1023 sequences for AF2, 8663 for RF2) with high diversity (Neff = 163.1 for AF2, 965.3 for RF2), whereas the 7YRW

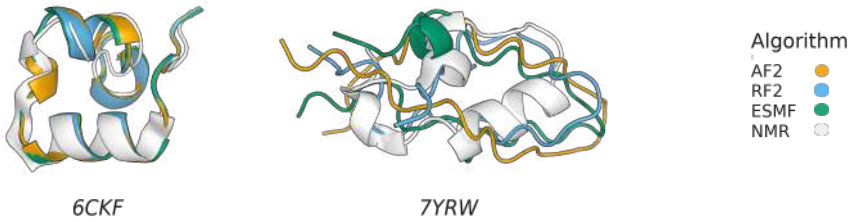
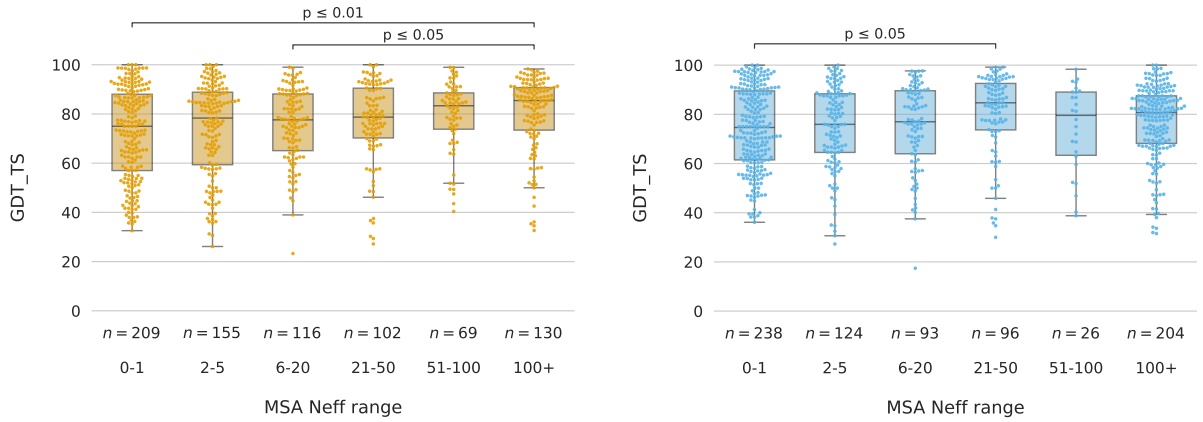


Fig. 4. Visualization of predicted structures for peptides 6CKF (left) and 7YRW (right), which share similar structural features but differ in prediction accuracy.

sequence resulted in a shallow alignment (4 sequences for AF2, 3 for RF2) with low diversity (Neff = 0.36 for both). The depth, diversity and coverage of the MSA allow the identification of coevolutionary information, providing robust spatial constraints that allowed for structural prediction of high quality by both AF2 and RF2. The shallower the MSA, the lower the algorithm's ability to extract reliable coevolutionary information. A quick analysis can reveal whether MSA are sufficiently rich and diverse to enable accurate structure prediction.



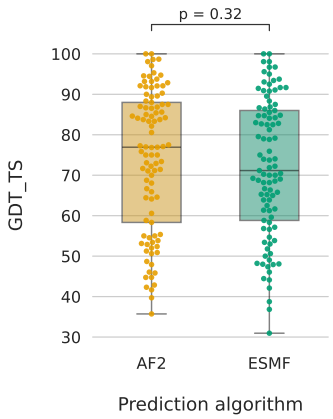
(a) GDT_TS score distribution for AF2 predictions across different MSA Neff ranges.

(b) GDT_TS score distribution for RF2 predictions across different MSA Neff ranges.

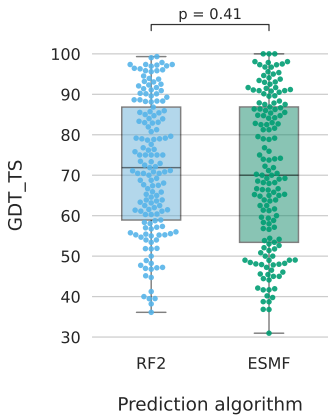
Fig. 5. Influence of MSA quality on structural prediction accuracy across methods. Significant differences ($p < 0.05$), assessed using Dunn's test with Bonferroni correction, are indicated on the top.

To better understand the impact of MSA quality on prediction accuracy, we analyzed the distribution of GDT_TS scores across different Neff ranges (Fig. 5). For AF2, significant differences were observed between the 0–1 and 100+ Neff ranges, as well as between the 6–20 and 100+ ranges (Dunn's test with Bonferroni correction, $p < 0.01$ and $p < 0.05$, respectively). For RF2, a significant difference was found between the 0–1 and 21–50 ranges ($p < 0.05$).

In cases where the MSA are particularly poor, sequence-based methods such as ESMF may be as effective as MSA-based methods to generate high-quality predictions. Specifically, for peptides where AF2 generated an empty MSA (depth = 1, *i.e.*, only the query sequence), the GDT_TS scores were not significantly different from those of ESMFold (Fig. 6a). Similar results were observed for RF2 (Fig. 6b). As an example, Fig. 4 also shows the predictions of ESMF for 6CKF (GDT_TS = 96.54) and 7YRW (GDT_TS = 57.26), which are similar to the scores obtained with the MSA-based methods AF2 and RF2.



(a) GDT_TS distribution for AF2 and ESMF predictions on peptides with empty MSA (AF2 input).



(b) GDT_TS distribution for RF2 and ESMF predictions on peptides with empty MSA (RF2 input).

Fig. 6. Comparison of prediction accuracy between MSA-based methods and ESMFold for peptides with empty MSA. Results of the Mann-Whitney U test are reported on the top.

Conclusion and future perspectives

Current methods, designed for protein structure prediction, perform slightly lower on peptide structure prediction, particularly for flexible peptides, whose conformational diversity remains challenging to capture. Our results hence show a range of performance spanning high quality peptide structure prediction (for peptides with protein-like structural features) and medium to low quality peptide structure prediction (for flexible peptides or predictions based on shallow MSA). Retraining these algorithms with high-quality peptide NMR structures [21] and additional NMR data could enhance accuracy and better capture the flexibility of peptide structures, particularly given the under-representation of these data in the current training datasets.

For MSA-based algorithms, the quality of the alignments remains a key factor in their performance. As peptides are shorter, it is essential to have MSA with high coverage, depth and diversity to allow the neural networks to extract maximum information on the spatial relationships between residues and subsequently make high-quality predictions. Adapting these alignments to shorter sequences would be a promising improvement option for these models.

Availability and Implementation

The dataset sequences in FASTA format and the scripts used for structure prediction, as well as the predicted model for each peptide from each tool can be downloaded at the following url: *https://zenodo.org/records/14887666*.

Acknowledgments

The authors are grateful for Dr. Pierre Tufféry for the insightful discussions.

Funding information

Clément Sauvestre is funded by ANRT CIFRE N°2022/0942.

References

[1] Muttenthaler M, King GF, Adams DJ, Alewood PF. Trends in peptide drug discovery. Nature Reviews Drug Discovery. 2021 Feb;20(4):309–325.

[2] Kryshtafovych A, Schwede T, Topf M, Fidelis K, Moult J. Critical assessment of methods of protein structure prediction (CASP)—Round XIV. Proteins: Structure, Function, and Bioinformatics. 2021;89(12):1607-17. Available from: https://onlinelibrary.wiley.com/doi/abs/10.1002/prot.26237.

[3] Jumper J, Evans R, Pritzel A, Green T, Figurnov M, Ronneberger O, et al. Highly accurate protein structure prediction with AlphaFold. nature. 2021;596(7873):583-9.

[4] Tunyasuvunakool K, Adler J, Wu Z, Green T, Zielinski M, Žídek A, et al. Highly accurate protein structure prediction for the human proteome. Nature. 2021;596(7873):590-6.

[5] McDonald EF, Jones T, Plate L, Meiler J, Gulsevin A. Benchmarking AlphaFold2 on peptide structure prediction. Structure. 2023;31(1):111-9.

[6] Baek M, Anishchenko I, Humphreys IR, Cong Q, Baker D, DiMaio F. Efficient and accurate prediction of protein structure using RoseTTAFold2. BioRxiv. 2023:2023-05.

[7] Rives A, Meier J, Sercu T, Goyal S, Lin Z, Liu J, et al. Biological Structure and Function Emerge from Scaling Unsupervised Learning to 250 Million Protein Sequences. PNAS. 2019. Available from: https://www.biorxiv.org/content/10.1101/622803v4.

[8] Lin Z, Akin H, Rao R, Hie B, Zhu Z, Lu W, et al. Language models of protein sequences at the scale of evolution enable accurate structure prediction. bioRxiv. 2022.

[9] Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, et al. The Protein Data Bank. Nucleic Acids Research. 2000 01;28(1):235-42. Available from: https://doi.org/10.1093/nar/28.1.235.

[10] Berman H, Henrick K, Nakamura H. Announcing the worldwide protein data bank. Nature structural & molecular biology. 2003;10(12):980-0.

[11] Steinegger M, Söding J. MMseqs2 enables sensitive protein sequence searching for the analysis of massive data sets. Nature biotechnology. 2017;35(11):1026-8.

[12] Kozma D, Simon I, Tusnády GE. PDBTM: Protein Data Bank of transmembrane proteins after 8 years. Nucleic Acids Research. 2012 11;41(D1):D524-9. Available from: https://doi.org/10.1093/nar/gks1169.

[13] White SH. Biophysical dissection of membrane proteins. Nature. 2009 May;459(7245):344-6. Available from: https://doi.org/10.1038/nature08142.

[14] Newport TD, Sansom MS, Stansfeld PJ. The MemProtMD database: a resource for membrane-embedded protein structures and their lipid interactions. Nucleic Acids Research. 2018 11;47(D1):D390-7. Available from: https://doi.org/10.1093/nar/gky1047.

[15] Lomize MA, Pogozheva ID, Joo H, Mosberg HI, Lomize AL. OPM database and PPM web server: resources for positioning of proteins in membranes. Nucleic Acids Research. 2011 09;40(D1):D370-6. Available from: https://doi.org/10.1093/nar/gkr703.

[16] Zemla A. LGA: a method for finding 3D similarities in protein structures. Nucleic acids research. 2003;31(13):3370-4.

[17] Morcos F, Pagnani A, Lunt B, Bertolino A, Marks DS, Sander C, et al. Direct-coupling analysis of residue coevolution captures native contacts across many protein families. Proceedings of the National Academy of Sciences. 2011;108(49):E1293-301. Available from: https://www.pnas.org/doi/abs/10.1073/pnas.1111471108.

[18] Simpkin AJ, Mesdaghi S, Sánchez Rodríguez F, Elliott L, Murphy DL, Kryshtafovych A, et al. Tertiary structure assessment at CASP15. Proteins: Structure, Function, and Bioinformatics. 2023;91(12):1616-35. Available from: https://onlinelibrary.wiley.com/doi/abs/10.1002/prot.26593.

[19] Wayment-Steele HK, Ojoawo A, Otten R, Apitz JM, Pitsawong W, Hömberger M, et al. Predicting multiple conformations via sequence clustering and AlphaFold2. Nature. 2024 Jan;625(7996):832-9. Available from: https://doi.org/10.1038/s41586-023-06832-9.

[20] del Alamo D, Sala D, Mchaourab HS, Meiler J. Sampling alternative conformational states of transporters and receptors with AlphaFold2. eLife. 2022 mar;11:e75751. Available from: https://doi.org/10.7554/eLife.75751.

[21] Timmons PB, Hewage CM. APPTEST is a novel protocol for the automatic prediction of peptide tertiary structures. Briefings in bioinformatics. 2021;22(6):bbab308.

RNA3DClust: Unsupervised segmentation of RNA 3D structures using density-based clustering

Quoc Khang LE[1], Eric ANGEL[1], Fariza TAHI[1], Guillaume POSTIC[1]

1 Université Paris-Saclay, Univ Evry, IBISC, 91020, Evry-Courcouronnes, France

Corresponding Author: quockhang.le@univ-evry.fr

Keywords:

RNA 3D structure, structural domains, clustering, lncRNAs

Abstract

A growing body of evidence shows that RNA function depends not only on primary and secondary structures but also on its 3D conformation. As the experimental determinations are costly and uncertain processes, computational prediction methods are essential. A critical task in such prediction is identifying substructures that can be modeled independently before assembling the global fold. In proteins, these are “structural domains” - yet no equivalent concept exists for RNA.

In this work, we present RNA3DClust, an adaptation of the Mean Shift algorithm for partitioning RNA 3D structures into compact, distinct regions, akin to protein domains. To evaluate the method, we built a reference dataset of annotated RNA 3D domains and developed a custom scoring scheme. We also show that RNA3DClust’s segmentations align with biologically and evolutionarily defined domains. Finally, with the emerging interest in long non-coding RNAs (lncRNAs), which likely contain folded substructures, we created a second dataset using predicted lncRNA models. RNA3DClust’s results on these models further demonstrate its potential for RNA domain analysis.

Introduction

Analysis involves breaking down a whole into parts to gain a better understanding. In structural biology, this means decomposing macromolecules into substructures. For proteins, these include: (i) secondary structures (local folds), (ii) supersecondary structures (assemblies of adjacent elements), and (iii) structural domains - independently folding or functional units [1]. Domains are essential to understanding protein function and are used in prediction tools like AlphaFold [2], which improves multimeric structure prediction by ~20% when domain boundaries are included [3,4]. RNA also follows a hierarchical structure: base pairing forms secondary structures; long-range interactions

form 3D structures, which are essential for biological roles such as catalysis, recognition, and regulation [5]. Despite of that, RNA 3D structural data remains scarce in databases like the PDB [6] and NAKB [7], and a formal notion of RNA structural domains is still missing. This is especially limiting for long non-coding RNAs (lncRNAs), which are biologically important and can span thousands of nucleotides, and thus, likely contain independently folded subregions similar to protein domains [8].

Here, we introduce the first computational study of RNA 3D domains, defined as compact, spatially distinct regions, inspired by Wetlaufer's definition for proteins [9]. We tested several clustering algorithms using RNA atomic coordinates to delineate domain boundaries. Due to the absence of reference datasets, we created our benchmark, using geometric and functional criteria. Our method, RNA3DClust, adapts the Mean Shift algorithm [10] for RNA-specific domain detection. While testing on a limited number of RNAs, RNA3DClust offers a basis for future RNA structure analysis.

Methods

We define RNA 3D domains as compact, spatially distinct regions [9], identified through the clustering of atomic coordinates. We retain only the C3' atom per nucleotide - part of the sugar-phosphate backbone and present in all residues. C3' is also used in alignment tools like RNA-align [11] and US-align [12], similar to the use of Cα [13] or Cβ [14] atoms in protein parsing.

To determine RNA 3D domains, a clustering algorithm for RNA 3D structures should meet three key criteria: (1) robustness to outlier, allowing detection of linker regions as outliers; (2) be able to identify clusters with irregular shapes, since RNA folds into diverse, non-globular conformations; and (3) non-parametric behavior, as the number of domains are unknown and varies across RNAs. We evaluated common clustering methods: k-means [15], hierarchical clustering [16], DBSCAN [17], Mean Shift [10], GMM [18], spectral clustering [19], and SOM [20] based on these criteria (Table 1).

Only DBSCAN and Mean Shift are the two methods that met all three criteria, as they define clusters based on point density rather than predefined shapes or counts. DBSCAN uses two hyperparameters: ε (neighborhood radius) and MinPts (minimum neighbors). A core point has at least MinPts within ε; clusters grow by linking core points and their neighbors. Points that are neither core nor reachable are marked as noise/outlier. Thus, selecting suitable ε and MinPts is crucial for effective clustering. Mean Shift locates dense regions by shifting each point toward the weighted center of its neighbors, using a kernel function. It depends on two hyperparameters: kernel type (uniform or Gaussian) and bandwidth h (search radius). Larger h merges clusters, smaller h resolves finer ones. We used scikit-learn's library [21] with a 10^{-3} convergence threshold and added Gaussian kernel support.

Since density-based clustering may yield spatially valid but scattered clusters along the RNA sequence, which is not biologically meaningful, we developed a post-clustering procedure to refine

results. It involves two main steps: processing outliers and processing labeled clusters. We defined eight rules (Fig. 1) and applied them iteratively until no changes occurred.

To assess RNA3DClust result, we used three metrics: Normalized Domain Overlap (NDO) [22], Domain Boundary Distance (DBD) [23], and our new index: Chain Segment Distance (CSD). Although have been widely used, both NDO and DBD have limitations. NDO can remain high even with the wrong number of domains, as it only reflects residue overlap. DBD, conversely, may be low despite near-correct boundary results due to strict thresholding [23]. It also assumes that only the reference has linkers and does not apply for single-domain cases, limiting its reliability in certain scenarios.

Therefore, we present CSD to offer a compromise between the NDO and DBD. Like the DBD, the scoring follows a decreasing pattern based on a distance threshold. However, rather than just boundaries, this distance is calculated between domains, like the NDO. For each computed domain i and true domain j, a score S_{ij} is calculated as:

$$S_{ij} = \frac{1}{2} \max \left\{ 2T - \left(d_{ij}^{5'} + d_{ij}^{3'} \right), 0 \right\}$$

where $d_{ij}^{5'}$ or $d_{ij}^{3'}$ is the distance (in nt) between the 5' or 3' end of the domain i and that of the domain j. The threshold T was set to T = 20 residues. Regarding linkers, their length is added to that of the adjacent domain, only if it lowers the value of $d_{ij}^{5'} + d_{ij}^{3'}$; otherwise, the length of the linker is not counted. Finally, the chain segment distance (CSD) score is calculated by this function:

$$\mathrm{CSD} = \begin{cases} \dfrac{\sum_{i=1}^{m} \max\limits_{1 \le j \le n} S_{ij}}{T \times m}, & \text{if } m \ge n \\ \dfrac{\sum_{j=1}^{n} \max\limits_{1 \le i \le m} S_{ij}}{T \times n}, & \text{otherwise} \end{cases}$$

where m and n are the total numbers of computed and true domains, respectively. The CSD score ranges from 0.0 to 1.0, with higher values indicating better segmentation. While developed for RNA domain evaluation, the CSD is more broadly applicable to other biological molecules.

Since RNA structural domains are not defined in the literature, we built our own two datasets by: (i) representative experimental RNA structures, and (ii) annotating domain boundaries using Wetlaufer's compactness and separation criteria [9]. For algorithm tuning, we created "Dataset 1" from experimental RNA structures in the PDB, and for benchmarking, we generated "Dataset 2" using predicted lncRNA 3D models. For Dataset 1, we extracted 4492 RNAs from the PDB. Domain annotation was semi-automated: initial Mean Shift clustering on C3' atoms followed by manual refinement in PyMOL. This yielded 163 two-domain and 4329 single-domain RNAs. To reduce redundancy, we selected 22 two-domain and 21 single-domain RNAs, resulting in 43 representative

entries. RNA under 30 nt were excluded. Given lncRNAs' emerging biological importance, Dataset 2 was generated with 69 predicted 3D lncRNA structures from LNCipedia [24], using RNAfold [25] for secondary structure prediction and RNAComposer [26,27] or AlphaFold3 [2] for 3D structures.

Results and Discussion

To evaluate the feasibility of hyperparameter tuning, we selected 8 arbitrary RNAs from Dataset 1: 4 single-domain (PDB: 3J3W, 5ZWN, 6ME0, 7BOI) and 4 two-domain (PDB: 3J29, 5XYI, 6NQB, 3JD5). For DBSCAN, we tested 1560 (ε, MinPts) combinations (ε: 5–30; MinPts: 1–60), but none of them gave the correct domain counts for more than 4 RNAs, indicating the need for finer and computationally intensive tuning. Thus, DBSCAN was excluded. For Mean Shift, we tested 2 kernel types (Gaussian, uniform) and 10 bandwidths (0.1–1.0), totaling 20 combinations. 4 settings yielded correct domain predictions for all 8 RNAs. Thus, we choose to refine the kernel type and bandwidth for Mean Shift.

For Mean Shift, larger bandwidths would yield fewer and larger clusters, therefore overly large bandwidths can bias results by producing single clusters for all single-domain RNAs. To mitigate this, we split Dataset 1 into 21 single-domain RNAs and 22 two-domain RNAs and tuned the two-domain RNAs subset. We observed that the results were more influenced by bandwidth than kernel type. No significant performance difference was observed between Gaussian and uniform kernels at the same bandwidth. A uniform kernel with 0.2 quantile bandwidth yielded the best scores: NDO = 0.806 ± 0.189, DBD = 0.308 ± 0.330, and CSD = 0.635 ± 0.297. Therefore, we selected a bandwidth = 0.2 quantile and uniform kernel as a default setting for RNA3DClust and used it for the rest of the article.

Across Dataset 1, RNA3DClust achieved average scores of NDO, DBD and CSD of 0.792 ± 0.173, 0.227 ± 0.350, and 0.700 ± 0.266, respectively.

To evaluate the strengths and limitations of RNA3DClust, we visualized in Fig.2 eight representative cases from Dataset 1: four correct (Fig. 2A-D) and four incorrect (Fig. 2E-H). For 4ADV and 6ME0 (Fig. 2A-B), RNA3DClust accurately segmented domains and removed outliers. For single-domain structure (2AAR - Fig. 2C), it identified one cluster with some outlier labeling. For 3JD5 (Fig. 2D), two domains were correctly found but slightly fragmented. These segmentations align well with assigned domains, and fragmented linkers may reflect actual domain flexibility.

The first two incorrect cases presented in Fig. 2 concern the same *O. cuniculus* 18S rRNA, captured at different stages of translation initiation: late (6YAN) and early (4KZZ). Interestingly, the mispredicted partitions are very different: one is over-segmented, the other is under-segmented (Figs. 3E and 3F, respectively), likely due to conformational changes during assembly of the pre-initiation complex. For 5ZWN (Fig. 2G), Mean Shift identified two clusters that post-processing failed to merge, causing

over-segmentation. For 5MY1 (Fig. 2H), both the clustering and post-clustering stages produced a two-domain partition inconsistent with the assigned domains.

In proteins, domains may also be defined by functional or evolutionary independence, which often aligns with 3D geometry. However, these criteria sometimes cause annotation inconsistencies [28]. Here, we examined how RNA3DClust's output compares with RNA functional domains in the literature. The first example is the 16S rRNA, known to consist of four functional domains: 5', Central, 3' Major, and 3' Minor [29] (Fig. 3A). Visual analysis of the *E. coli* 16S rRNA (PDB: 3J29) led us to annotate two structural domains - one corresponding to the 3' Major, and the other overlapping the remaining functional domains (Fig. 3B). RNA3DClust reproduced this two-domain partitioning, with one cluster aligning well with the 3' Major functional domain (Fig. 3C).

Rfam [30] annotates "Domain I" of the *Oceanobacillus iheyensis* group II intron RNA as "group-II-D1D4-3" based on conserved sequence and secondary structure (Fig. 3D). In our reference (PDB: 4Y1N), this RNA is split into 2 structural domains, one of which overlaps that domain (Fig. 3E). RNA3DClust correctly identified this region, producing a cluster spanning residues 94–232 (Fig. 3F). This agreement between structural and evolutionary definitions highlights RNA3DClust's ability to detect biologically meaningful RNA substructures.

We then applied RNA3DClust to the 69 lncRNA 3D models in Dataset 2, obtaining average scores of NDO, DBD and CSD values of 0.617 ± 0.206, 0.379 ± 0.290, and 0.354 ± 0.294, respectively. These values are generally lower than those for Dataset 1, likely due to Dataset 2's non-native, predicted structures, which may lack of proper hierarchical organization.

Based on those scores, we selected 4 correct cases (Fig. 4A-D) and 4 with limitations (Fig. 4E-H). In the correct ones, post-clustering improved continuity by extending clusters over outliers. RNA3DClust successfully detected the compact and helical domains in AADACL2 (Fig. 4A) and ADGRL3-AS1:5 (Fig. 4B). RERG-AS1:1 (Fig. 4C) was well segmented after extending over an outlier and removing a redundant cluster. For LINC01016:6 (Fig. 4D), the compact domain was identified, with partial detection of the helical region.

In this study, helices were treated as domains, though it's hard to separate them from unfolded linkers. Therefore, segmentation struggled with helix-rich RNAs. In Fig. 4E, most of the helical domain was mislabeled as outliers. In Fig. 4F-G, multi-domain RNAs were under-segmented. Conversely, Fig. 4H shows an over-segmented two-helix RNA, split into four clusters and outliers.

In conclusion, RNA3DClust performed well across both datasets, though it still has some limitations when dealing with the structural diversity in Dataset 2. The Mean Shift algorithm with adaptations also shows promise for partitioning biological structures like RNAs and potentially proteins.

	k-means	Hierarchical	DBSCAN	Mean Shift	GMM	Spectral	SOM
Robust against noise		✔	✔	✔		✔	✔
Find irregular shapes		✔	✔	✔	✔	✔	✔
Non-parametric			✔	✔			

Tab. 1: *Comparison between widely-used clustering algorithms for partitioning RNA 3D structures.*

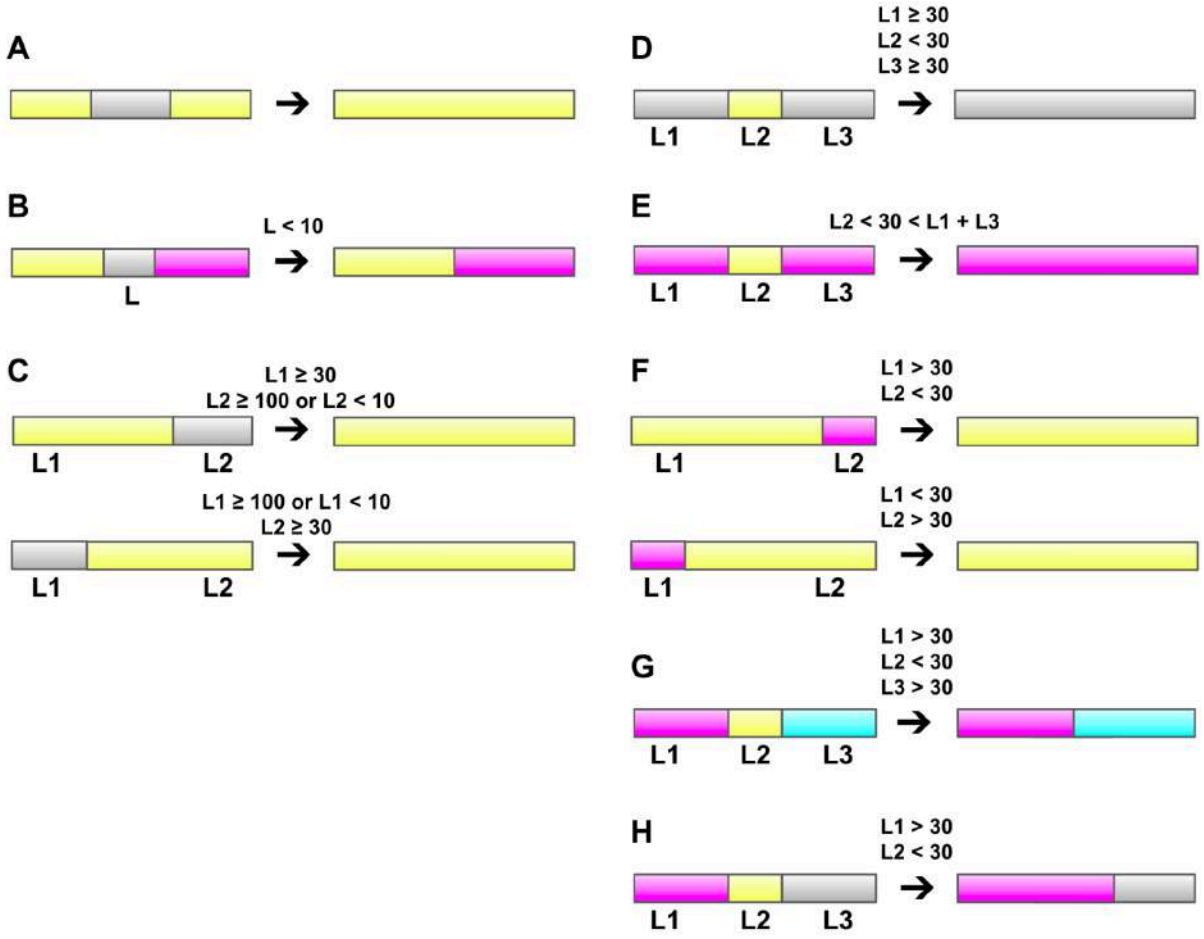


Fig. 1: *The eight rules for the post-clustering procedure. (**A, B**, and **C**) The rules for outliers; (**D, E, F, G**, and **H**) The rules for labeled clusters. The gray color indicates outlier regions. The yellow, magenta and cyan colors indicate cluster regions. The sequence length of the left, middle and right regions are symbolized by "L1", "L2" and "L3", respectively. On each panel, the segments on the left represent the clustering result before the post-clustering procedure, the segments on the right represent the clustering result after the post-clustering procedure.*

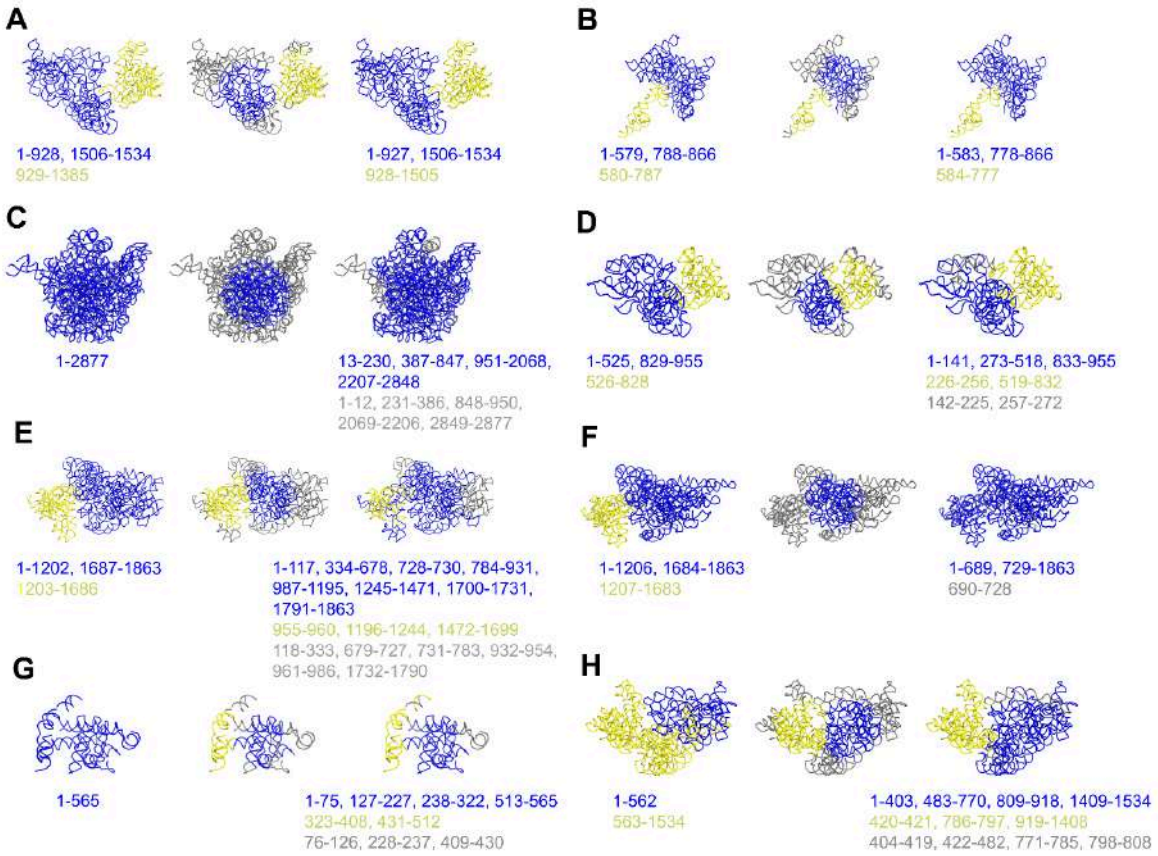


Fig. 2: *Examples of segmentations for experimental RNA 3D structures (Dataset 1). For each RNA, the reference segmentation is presented on the left panel; on the middle and right panels are RNA3DClust results, before and after the post-clustering procedure, respectively. Labeled clusters are colored in blue and yellow, while outliers are in gray. Domain positions are below the structures. (**A**) E. coli 16S rRNA (PDB entry: 4ADV), (**B**) T. vestitus Th.e.I3 group II intron RNA (6ME0), (**C**) D. radiodurans 23S rRNA (2AAR), (**D**) B. taurus mitochondrial 28S rRNA (3JD5), (**E**) O. cuniculus 18S rRNA, part of the 80S initiation complex (6YAN), (**F**) O. cuniculus 18S rRNA, part of the 48S pre-Initiation complex (4KZZ), (**G**) S. cerevisiae U1 spliceosomal RNA (5ZWN), (**H**) E. coli 16S rRNA (5MY1).*

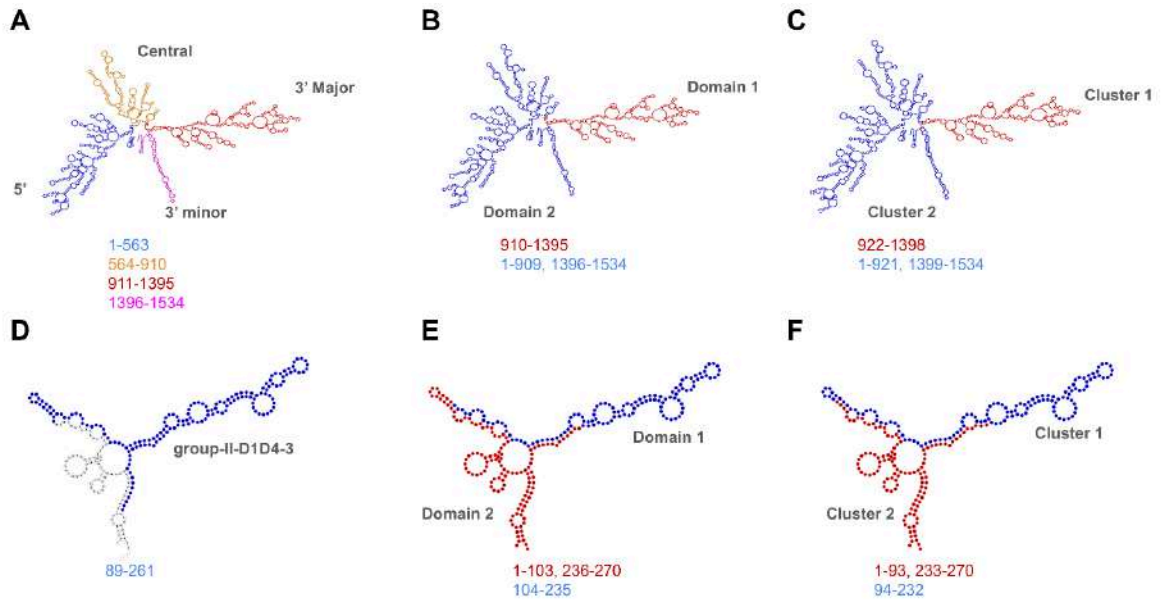


Fig. 3: *Comparison between domain annotations based on (**A**) biological function, (**D**) evolution, (**B**, **E**) RNA 3D structure, and (**C**, **F**) the clustering performed by RNA3DClust. Two examples are shown: (**A**, **B**, **C**) the E. coli 16S rRNA (PDB entry: 3J29), and (**D**, **E**, **F**) the O. iheyensis group II intron domain I RNA (4Y1N). The labels "Central", "5'", "3' Major", and "3' minor" are the names given to the functional domains in the literature. The label "group-II-D1D4-3" is the name of the RF02001 family in Rfam.*

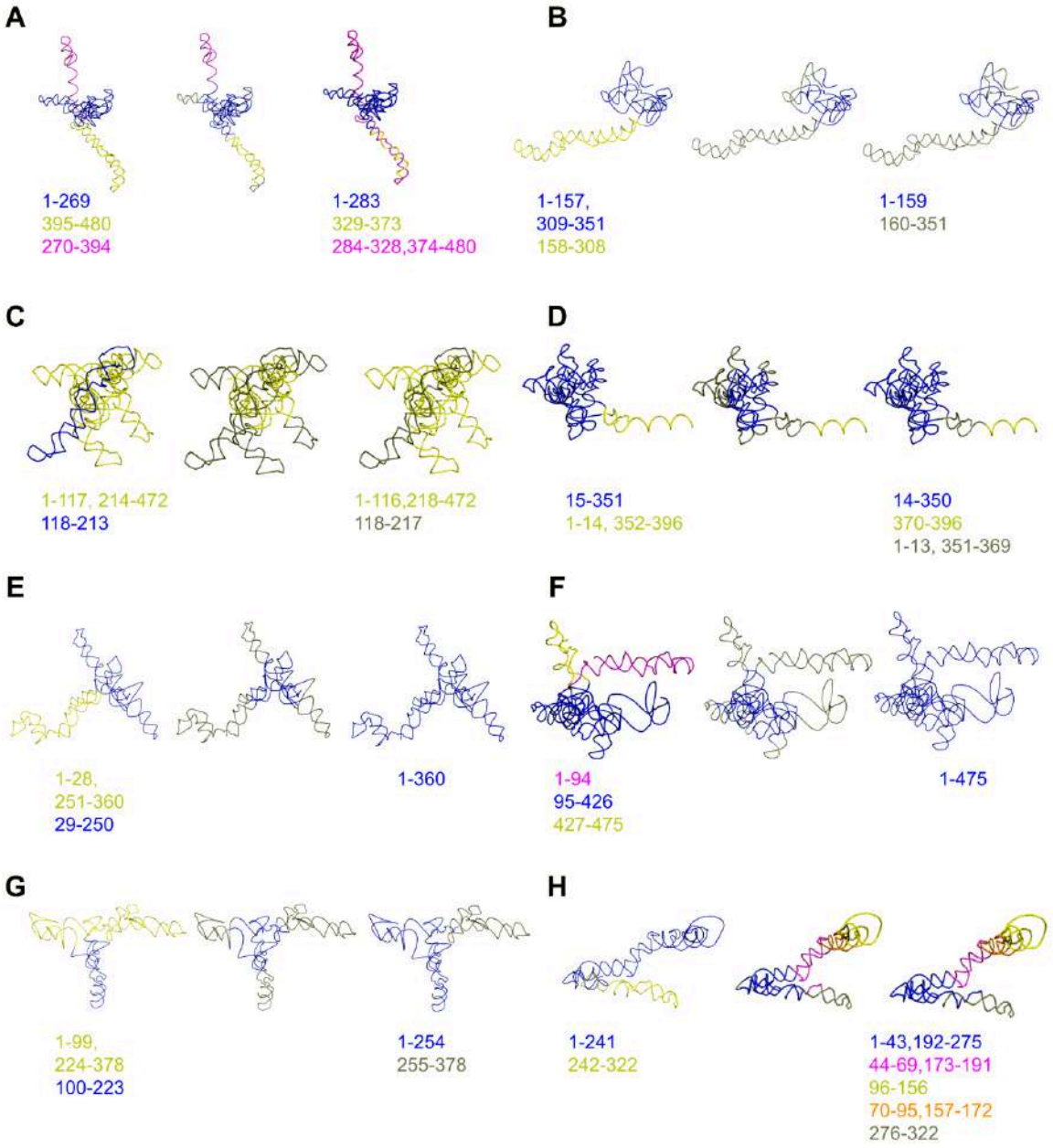


Fig. 4: *Examples of segmentations for predicted 3D structures of lncRNAs (Dataset 2). For each RNA, the reference segmentation is presented on the left panel; on the middle and right panels are RNA3DClust results, before and after the post-clustering procedure, respectively. Labeled clusters are colored in blue, yellow, orange and magenta, while outliers are in gray. Domain positions are below the structures. The LNCipedia entries are: (**A**) AADACL2-AS1:1; (**B**) ADGRL3-AS1:5; (**C**) RERG-AS1:1; (**D**) LINC01016:6; (**E**) ADORA2A-AS1:16; (**F**) ADGRA1-AS1:2; (**G**) AADACL2-AS1:6; and (**H**) lnc-AADACL2-1:5.*

References

1. Richardson JS. The Anatomy and Taxonomy of Protein Structure [Internet]. In: Anfinsen CB, Edsall JT, Richards FM, editors. Advances in Protein Chemistry. Academic Press; 1981 [cited 2025 Jan 8]. page 167–339.Available from: https://www.sciencedirect.com/science/article/pii/S0065323308605203

2. Abramson J, Adler J, Dunger J, Evans R, Green T, Pritzel A, et al. Accurate structure prediction of biomolecular interactions with AlphaFold 3. Nature 2024;630:493–500.

3. Evans R, O'Neill M, Pritzel A, Antropova N, Senior A, Green T, et al. Protein complex prediction with AlphaFold-Multimer [Internet]. 2022 [cited 2024 Nov 7];2021.10.04.463034. Available from: https://www.biorxiv.org/content/10.1101/2021.10.04.463034v2

4. Bret H, Gao J, Zea DJ, Andreani J, Guerois R. From interaction networks to interfaces, scanning intrinsically disordered regions using AlphaFold2. Nat. Commun. 2024;15:597.

5. Micura R, Höbartner C. Fundamental studies of functional nucleic acids: aptamers, riboswitches, ribozymes and DNAzymes. Chem. Soc. Rev. 2020;49:7331–53.

6. wwPDB consortium. Protein Data Bank: the single global archive for 3D macromolecular structure data. Nucleic Acids Res. 2019;47:D520–8.

7. Lawson CL, Berman HM, Chen L, Vallat B, Zirbel CL. The Nucleic Acid Knowledgebase: a new portal for 3D structural information about nucleic acids. Nucleic Acids Res. 2023;gkad957.

8. Arunkumar G. LncRNAs: the good, the bad, and the unknown. Biochem. Cell Biol. Biochim. Biol. Cell. 2024;102:9–27.

9. Wetlaufer DB. Nucleation, Rapid Folding, and Globular Intrachain Regions in Proteins. Proc. Natl. Acad. Sci. U. S. A. 1973;70:697–701.

10. Fukunaga K, Hostetler L. The estimation of the gradient of a density function, with applications in pattern recognition. IEEE Trans. Inf. Theory 1975;21:32–40.

11. Gong S, Zhang C, Zhang Y. RNA-align: quick and accurate alignment of RNA 3D structures based on size-independent TM-scoreRNA. Bioinformatics 2019;35:4459–61.

12. Zhang C, Shine M, Pyle AM, Zhang Y. US-align: universal structure alignments of proteins, nucleic acids, and macromolecular complexes. Nat. Methods 2022;19:1109–15.

13. Wells J, Hawkins-Hooker A, Bordin N, Sillitoe I, Paige B, Orengo C. Chainsaw: protein domain segmentation with fully convolutional neural networks. Bioinformatics 2024;40:btae296.

14. Zhu K, Su H, Peng Z, Yang J. A unified approach to protein domain parsing with inter-residue distance matrix. Bioinformatics 2023;39:btad070.

15. Lloyd S. Least squares quantization in PCM. IEEE Trans. Inf. Theory 1982;28:129–37.

16. Bridges CC. Hierarchical Cluster Analysis. Psychol. Rep. 1966;18:851–4.

17. Ester M, Kriegel HP, Sander J, Xu X. A density-based algorithm for discovering clusters in large spatial databases with noise. In: Proceedings of the Second International Conference on Knowledge Discovery and Data Mining. Portland, Oregon: AAAI Press; 1996. page 226–31.

18. Dempster AP, Laird NM, Rubin DB. Maximum Likelihood from Incomplete Data Via the EM Algorithm. J. R. Stat. Soc. Ser. B Methodol. 1977;39:1–22.

19. Donath WE, Hoffman AJ. Lower Bounds for the Partitioning of Graphs. IBM J. Res. Dev. 1973;17:420–5.

20. Kohonen T. Self-organized formation of topologically correct feature maps. Biol. Cybern. 1982;43:59–69.

21. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, et al. Scikit-learn: Machine Learning in Python. J. Mach. Learn. Res. 2011;12:2825–30.

22. Tai CH, Lee WJ, Vincent JJ, Lee B. Evaluation of domain prediction in CASP6. Proteins 2005;61 Suppl 7:183–92.

23. Tress M, Cheng J, Baldi P, Joo K, Lee J, Seo JH, et al. Assessment of predictions submitted for the CASP7 domain prediction category. Proteins 2007;69 Suppl 8:137–51.

24. Wang L, Zhong H, Xue Z, Wang Y. Res-Dom: predicting protein domain boundary from sequence using deep residual network and Bi-LSTM. Bioinforma. Adv. 2022;2:vbac060.

25. Xue Z, Xu D, Wang Y, Zhang Y. ThreaDom: extracting protein domain boundary information from multiple threading alignments. Bioinformatics 2013;29:i247–56.

26. Shi Q, Chen W, Huang S, Jin F, Dong Y, Wang Y, et al. DNN-Dom: predicting protein domain boundary from sequence alone by deep neural network. Bioinforma. Oxf. Engl. 2019;35:5128–36.

27. Volders PJ, Anckaert J, Verheggen K, Nuytens J, Martens L, Mestdagh P, et al. LNCipedia 5: towards a reference set of human long non-coding RNAs. Nucleic Acids Res. 2019;47:D135–9.

28. Lorenz R, Bernhart SH, Höner zu Siederdissen C, Tafer H, Flamm C, Stadler PF, et al. ViennaRNA Package 2.0. Algorithms Mol. Biol. 2011;6:26.

29. Sarzynska J, Popenda M, Antczak M, Szachniuk M. RNA tertiary structure prediction using RNAComposer in CASP15. Proteins 2023;91:1790–9.

30. Popenda M, Szachniuk M, Antczak M, Purzycka KJ, Lukasiak P, Bartol N, et al. Automated 3D structure composition for large RNAs. Nucleic Acids Res. 2012;40:e112.

31. Bernard C, Postic G, Ghannay S, Tahi F. Has AlphaFold3 achieved success for RNA? Acta Crystallogr. Sect. Struct. Biol. 2025;81:49–62.

32. Zheng W, Zhou X, Wuyun Q, Pearce R, Li Y, Zhang Y. FUpred: detecting protein domains through deep-learning-based contact map prediction. Bioinformatics 2020;36:3749–57.

33. Postic G, Ghouzam Y, Chebrek R, Gelly JC. An ambiguity principle for assigning protein structural domains. Sci. Adv. 2017;3:e1600552.

34. Xu Z, Culver GM. Differential assembly of 16S rRNA domains during 30S subunit formation. RNA 2010;16:1990–2001.

35. Ontiveros-Palacios N, Cooke E, Nawrocki EP, Triebel S, Marz M, Rivas E, et al. Rfam 15: RNA families database in 2025. Nucleic Acids Res. 2024;53:D258–67.

36. Lau AM, Kandathil SM, Jones DT. Merizo: a rapid and accurate protein domain segmentation method using invariant point attention. Nat. Commun. 2023;14:8445.

37. Eguchi RR, Huang PS. Multi-scale structural analysis of proteins by deep semantic segmentation. Bioinformatics 2020;36:1740–9.

Session 5: Evolution, phylogeny and comparative genomics

Natural selection acting on gene expression and regulation in molerats

Maeële DAUNESSE[,2], Elise PAREY, Diego VILLAR, Camille BERTHELOT[,5]

1 Institut Pasteur, Université de Paris, CNRS UMR 3525, INSERM UA12, Comparative Functional Genomics group, F-75015 Paris, France.

2 EMBL-EBI, Wellcome Genome Campus, Hinxton, UK

3 Centre for Life's Origins & Evolution, Dept of Genetics, Evolution & Environment, University College London, London, UK

4 Blizard Institute, Barts and the London School of Medicine and Dentistry, Queen Mary University of London, 4 Newark Street, London E1 2AT

5 Institut de Biologie de l'Ecole Normale Supérieure (IBENS), Ecole Normale Supérieure, CNRS, INSERM, Univ. PSL, Paris, France

Corresponding Author: maelle.daunesse@pasteur.fr

Keywords

Comparative genomics, Mammalian evolution, Gene expression, Gene regulation, Phylogenetic

Abstract

Understanding the genetic basis of phenotypic adaptations poses a significant challenge in evolutionary genomics. Despite the morphological and physiological diversity in mammalian traits, their coding genomes exhibit a high degree of conservation, implying that changes in gene expression and regulation are pivotal in driving phenotype evolution. This study aims to identify shifts in gene expression and cis-regulatory activity and their potential role in phenotypic adaptation. Using African mole-rats as a model, renowned for their unique phenotypic adaptation traits like cancer resistance and hypoxia tolerance, we aimed to elucidate the genome-wide gene expression patterns underlying these traits that have been mainly characterised at the level of candidate genes and in individual species. Profiling gene expression in heart and liver tissues across two mole-rat species and two rodent outgroups, we used a phylogenetic comparative approach to identify genes with expression shifts within the mole-rat clade and in specific genera. These shifted genes are associated with functions pertinent to known adaptations in naked mole-rats, such as cellular respiration and glycolysis in the heart. Furthermore, our analysis revealed concordant changes in the regulatory landscape of these genes. By employing a phylogenetic comparative approach, we offer new insights into the interplay between gene expression, regulation, and phenotypic evolution in mammals. Our findings shed light on the molecular mechanisms driving the evolution of unique traits in mole-rats and potentially other mammalian species.

Evolutionary dynamics of centromeric DNA in guenon might end an old anthropocentric dogma

Julien PICHON1, Lauriane CACHEUX1,2, Manel AIT EL HADJ[1], Axel JENSEN[3], Katerina GUSCHANSKI[3,4], Loïc PONGER[1] and Christophe ESCUDÉ[1]

1 Museum National d'Histoire Naturelle, Structure et instabilité des génomes, UMR7196, Paris 75231, France

2 Museum National d'Histoire Naturelle, Institut de Systématique, Évolution, Biodiversité (ISYEB), UMR7205, Paris 75231, France

3 Institute of Ecology and Evolution, School of Biological Sciences, University of Edinburgh, Edinburgh EH8 9XP, UK

4 Department of Ecology and Genetics, Animal Ecology, Uppsala University, SE-75236 Uppsala, Sweden

Corresponding Author: christophe.escude@mnhn.fr

Keywords

Alpha-Satellite DNA, Centromere, Evolution, Primates

Abstract

The characterization of centromeric DNA has recently seen a major breakthrough thanks to advances in sequencing technologies, reaching complete resolution for several human cell lines and other primates through Telomere-to-Telomere (T2T) assemblies. New bioinformatic tools have been developed to describe the main DNA type composing the centromere in most Primates, alpha satellite (AS) DNA. However, both the tools and the resulting annotations often rely on the human genome to interpret the evolutionary dynamics of other species. One example is the use of human AS families to annotate AS sequences in other species, despite the rapid evolution of these sequences. Moreover, the current model of alpha satellite evolution, which is primarily based on observations in humans, needs to be tested against data from other primates. As T2T sequencing remains costly and technically demanding, we propose an alternative approach that directly leverages long-read sequencing data. In the present study, we identified AS-containing reads within a Pacbio HiFi dataset for a Cercopithecini species, *Cercopithecus cephus*. Through a de novo annotation of these sequences, we identified two families that we previously detected in two related species, as well as a new family, which is the least abundant but also the most ancient. These three families also appear to be spatially segmented across the genome, corresponding to distinct evolutionary layers. To investigate the organization of alpha satellite monomers within these layers, we developed a tool designed to detect higher-order repeat (HOR) structures without relying on predefined family classifications. Unlike humans, *C. cephus* exhibits a predominantly monomeric-like organization of its AS, with only 1.6% of sequences forming HORs. Interestingly, these HORs are mainly found in the oldest evolutionary layers, suggesting a potential transition from HOR to a monomeric organization in this species. These findings support the idea that HOR organization is not a unique or highly specialized structure and could arise independently in multiple clades.

A Comprehensive Study of Inverted Repeats in Prokaryotic Genomes: Enrichment, Depletion, and Taxonomic Variations

Victor BANON GARCIA[1], Ivan JUNIER[1] and Nelle VAROQUAUX[1]

1 TIMC, Université Grenoble Alpes, CNRS, Grenoble INP, Grenoble 38000, France

Corresponding author: nelle.varoquaux@univ-grenoble-alpes.fr,ivan.junier@univ-grenoble-alpes.fr

Keywords

Inverted repeat, prokaryotes, pattern finding, systematic analysis

Abstract

Inverted repeats (IRs) are genetic elements with a DNA motif (left arm) followed by a gap, or spacer, and its reverse complement (right arm) – *e.g.*, ATACGGnnnCCGTAT. They play a key role in many biological functions, including gene regulation, DNA replication, and genome plasticity. In this study, we aim to systematically investigate the distribution of short IRs (with gap lengths up to 20 bp) in all completely sequenced prokaryotic species by confronting observed statistics to expected ones computed by permuting DNA sequences under specific constraints. Through this systematic approach, we reveal complex patterns of IR biases with five main observations: (i) a systematic enrichment of IRs with arm lengths longer than 6 bp, (ii) a systematic depletion of palindromes (IRs with a zero-length gap) shorter than 6 bp, (iii) qualitatively different biases between coding and non-coding regions, (iv) in non-coding regions, the most frequent enrichment over bacterial species occurs for gap length of 4 bp similar to the most common loop size of RNA hairpins associated with known transcription terminators in model organisms, and (v) biases in coding regions that strongly depend on the species considered. Altogether, these findings — both corroborating and further deepening previous analyses — highlight universal evolutionary constraints as well as species-specific selective pressures that act on genome sequences, particularly on IRs.

Session 6: Functional and Integrative Genomics

rnaends: an R package targeted to study the exact RNA ends at the nucleotide resolution

Tomas CAETANO[1], Peter REDDER[1], Gwennaele FICHANT[1], Roland BARRIOT[*1]

1 Laboratoire de Microbiologie et Génétique Moléculaires (LMGM), Centre de Biologie Intégrative (CBI), Université de Toulouse, CNRS, Toulouse, France.

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

Keywords

RNA-end sequencing, 5' end, 3' end, EMOTE, R-package, Transcription start site, endoribonucleolytic cleavage site, RNA degradation, p

Abstract

5' and 3' RNA-end sequencing protocols have unlocked new opportunities to study aspects of RNA metabolism such as synthesis, maturation and degradation, by enabling the quantification of exact ends of RNA molecules in vivo. From RNA-Seq data that have been generated with one of the specialized protocols, it is possible to identify transcription start sites (TSS) and/or endoribonucleolytic cleavage sites, and even co-translational 5' to 3' degradation dynamics in some cases. Furthermore, post-transcriptional addition of ribonucleotides at the 3' end of RNA can be studied at the nucleotide resolution.

While different RNA-end sequencing library protocols can vary, and each have their specificities, the generated RNA-Seq data are very similar and share common processing steps. Most importantly, the major aspect of RNA-end sequencing is that only the 5' or 3' end mapped location is of interest, contrary to conventional RNA sequencing that considers genomic ranges for gene expression analysis. This translates to a simple representation of the quantitative data as a count matrix of RNA-end location on the reference sequences. This representation seems under-exploited and is, to our knowledge, not available in a generic package focused on the analyses on the exact transcriptome ends.

Here, we present the rnaends R package which is dedicated to RNA-end sequencing analysis. It offers features for raw read pre-processing, RNA-ends mapping and quantification, RNA-ends count matrix post-processing, and further count matrix downstream analyses such as TSS identification, fast Fourier transform for signal periodic patterns analysis, or differential proportion of RNA-ends analysis. The use of rnaends is illustrated with applications in RNA metabolism studies through selected workflows on published RNA-end datasets: (i) TSS identification, (ii) ribosome translation speed and co-translational degradation, (iii) post-transcriptional modifications analysis and differential proportion analysis.

Benchmarking circRNA Detection Tools from Long-Read Sequencing Using Data-Driven and Flexible Simulation Framework

Anastasia RUSAKOVICH[1], Sebastien CORRE[1], Edouard CADIEU[1], Rose-Marie FRABOULET[1], Marie-Dominique GALIBERT [1,2], Thomas DERRIEN[1*] and Yuna BLUM[1*]

1 Univ Rennes, CNRS, INSERM, IGDR (Institut de Génétique et Développement de Rennes) - UMR 6290, 2 Av. du Professeur Léon Bernard Bâtiment 4, F-35000 Rennes, France.

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

*Co-last Author

Corresponding Authors: anastasia.rusakovich@univ-rennes.fr, thomas.derrien@univ-rennes.fr, yuna.blum@univ-rennes.fr

Keywords

circRNA, long-read nanopore sequencing, benchmark comparison, simulation framework

Abstract

Circular RNAs (circRNAs) are unique non-coding RNAs with covalently closed loop structures formed through backsplicing events. Their stability, tissue-specific expression patterns, and potential as disease biomarkers have garnered increasing attention. However, their circular structure and diverse size range pose challenges for conventional sequencing technologies. Long-read Oxford Nanopore (ONT) sequencing offers promising capabilities for capturing entire circRNA molecules without fragmentation, yet the effectiveness of bioinformatic tools for analyzing this data remains understudied.

This study presents the first benchmark comparison of three specialized tools for circRNA detection from ONT long-read data: CIRI-long [1], IsoCIRC [2], and circNICK-Irs [3]. To address the lack of standardized evaluation frameworks, we developed a novel computational pipeline, open-source and freely available, to generate realistic simulated circRNA ONT long-read datasets. Our pipeline integrates several molecular features of circRNAs extracted from established databases - circAtlas [4] and circBase [5] and real datasets into NanoSim tool [6] and outputs FASTQ reads reflecting therefore biological diversity and technical properties.

We assessed tool performance across key metrics, including precision, recall, specificity, accuracy, and F1 score. Our analysis revealed distinct performance profiles: while all tools exhibited high specificity, they varied in precision and their ability to detect different circRNA subtypes, often showing limited sensitivity and precision. Notably, the overlap in detected circRNAs among tools was relatively low. Additionally, computational efficiency varied significantly across the tools. This suggests that relying

on a single tool might not be ideal, and combining tools or improving algorithms could be necessary for more accurate circRNA detection from ONT data.

This benchmark provides valuable insights for researchers selecting appropriate tools for circRNA studies using ONT sequencing. Furthermore, our customizable simulation framework, offering a resource to optimize detection approaches and advance bioinformatic tool development for circRNA research is freely available at: https://gitlab.com/bioinfog/circall/nano-circ.

Introduction

Non-coding RNAs (ncRNAs), which make up a significant portion of the transcriptome, have emerged as critical players in various cellular processes, ranging from gene regulation to disease pathogenesis [7]. Among them, circular RNAs (circRNAs) represent a unique class of ncRNAs, formed through non-canonical backsplicing events, resulting in a covalently closed loop structure. Since their discovery, circRNAs have garnered increasing attention due to their stability, tissue-specific expression, and potential roles as biomarkers in diseases, including cancer [8,9]. However, their circular structure and diverse size range—spanning from less than 100 to almost 100000 nucleotides—pose significant challenges to conventional sequencing technologies, particularly second-generation sequencing methods, which often rely on read fragmentation.

Recent advancements in third-generation sequencing technologies, such as Oxford Nanopore sequencing, have provided novel opportunities to explore circRNAs in greater detail [10]. Unlike second-generation methods, Nanopore sequencing offers long-read capabilities that can capture entire circRNA molecules without the need for fragmentation, making it a promising approach for the comprehensive characterization of circRNAs. However, the full potential of long-read sequencing for circRNA discovery and annotation depends on bioinformatics tools that can accurately detect and quantify circRNAs from these datasets.

Despite the growing number of bioinformatics tools developed for circRNA detection [11], a comprehensive benchmark comparing their performance on long-read Nanopore sequencing data remains absent. To address this gap, we performed a systematic comparison of three circRNA detection tools — CIRI-long[1], IsoCIRC[2] and circNICK-lrs[3]. We selected these tools because they represent the three major methodological approaches to long-read circRNA detection, each with distinct experimental protocols and computational pipelines: rolling circle reverse transcription (CIRI-long), rolling circle amplification followed by nanopore sequencing (isoCirc), and direct circRNA

linearization approaches (circNick-LRS). This selection provides comprehensive coverage of the current landscape of long-read circRNA detection methodologies, enabling evaluation of how different experimental and computational strategies affect detection performance across various circRNA characteristics.

The complexity of circRNA detection requires a robust and reproducible evaluation framework that goes beyond traditional wet-lab datasets. Experimental data inherently suffers from limitations such as biological variability, sequencing biases, and the challenge of definitively establishing ground truth. Simulated datasets offer a solution to these challenges, providing a controlled environment with precisely known circRNA and linear RNA annotations. By generating *in silico reads* that mimic the molecular and sequencing characteristics of Nanopore technologies, we can create a comprehensive ground truth dataset that allows for assessment of performance by each of circRNA detection tools.

In this study, we provide a first framework for the generation of simulation data of long-read sequencing of circRNA and present the first benchmark analysis of these 3 tools using simulated datasets. By evaluating their performance across different metrics, we aim to provide valuable insights into the strengths and limitations of each tool, guiding researchers in selecting the most appropriate software for their circRNA studies (Fig.1).

Material & Methods

Benchmarking framework

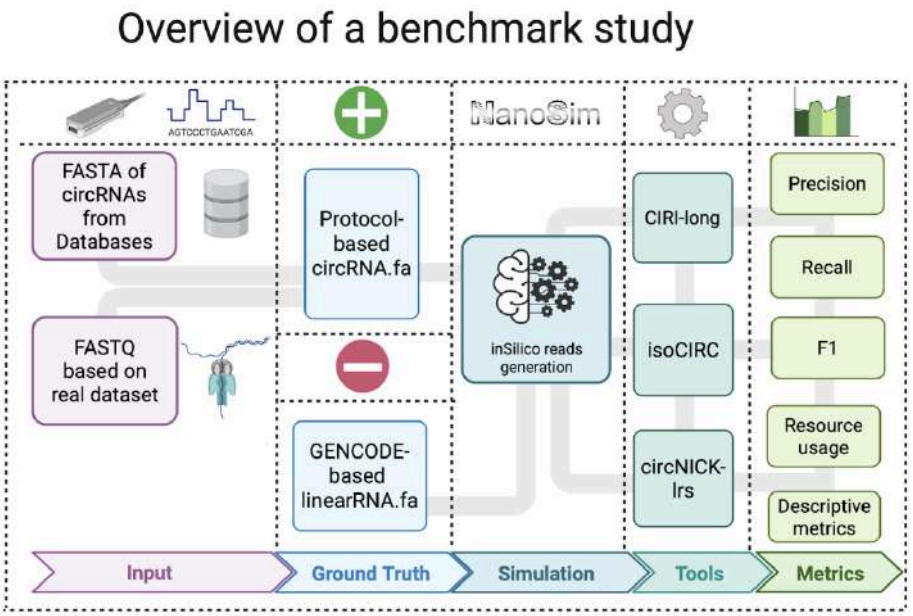


Fig. 1: Overview of the benchmarking study. *Schematic representation of the benchmarking framework for circular RNA (circRNA) detection tools using mouse brain nanopore sequencing data as the training model. The workflow progresses through five stages: (I) Input data preparation including FASTA sequences from circRNA databases and FASTQ files from real wet-lab protocols for training; (II) Ground truth preparation from simulation parameters; (III) In silico read simulation using NanoSim version 3.1.0; (IV) Computational analysis using three long-read circRNA detection tools: CIRI-long, isoCirc, and circNick-LRS; (V) Comprehensive performance evaluation through multiple metrics including sensitivity, precision, specificity, accuracy, and F1 score, calculated across different overlap thresholds.*

Tools for circular RNA identification using long-read sequencing data

CIRI-long [1]: is a computational method designed for profiling circRNAs using nanopore long-read sequencing data based on a rolling circle reverse transcription approach. The algorithm reconstructs full-length circRNA sequences through a multi-step process: (1) **Data preprocessing**: demultiplexing, quality control, and adapter removal; (2) **Repetitive pattern identification**: k-mer-based detection of circular patterns using k=8 and k=11 with homopolymer-compressed k-mers; (3) **Consensus sequence generation**: partial order alignment (SPOA) to create cyclic consensus sequences with 80% similarity threshold between repetitive segments; (4) **Mapping and BSJ detection**: alignment using minimap2 for sequences >150 bp and bwa mem [12] for shorter sequences, followed by iterative alignment strategy; (5) **Filtering and validation**: canonical splice signal detection (GT/AG, GC/AG, AT/AC) and clustering based on genomic coordinates. The method validates circRNAs against the circAtlas database. More details can be found at [https://github.com/bioinfo-biols/CIRI-long].

isoCirc [2]: is a method for sequencing and characterizing full-length circular RNA isoforms using rolling circle amplification followed by nanopore long-read sequencing. The computational pipeline involves: (1) **Data preprocessing**: demultiplexing, quality control, and adapter removal; (2) **Repetitive pattern identification**: tandem repeat detection to identify multiple copies of circRNA sequences within reads; (3) **Consensus sequence generation**: construction of consensus sequences from detected tandem repeats with copy number-dependent error correction (average copy number of 14.5); (4) **Mapping and BSJ detection**: alignment of consensus sequences to reference genome using minimap2, with BSJ identification through split-read analysis; (5) **Filtering and validation**: multi-tiered alignment scoring, stringent validation of back-spliced junctions (BSJs) and forward-spliced junctions (FSJs), requiring high mapping quality and fidelity. The method characterizes alternative splicing

events and validates against circBase and MiOncoCirc databases. More information can be found at [https://github.com/Xinglab/isoCirc].

circNick-LRS [3]: is a computational workflow for profiling circRNAs from linearized circRNA nanopore long-read sequencing data. The pipeline consists of: (1) **Data preprocessing**: demultiplexing, quality control and length filtering; (2) **Mapping**: direct alignment to reference genome using pblat with parallelized implementation; (3) **BSJ detection**: identification of back-splice junctions through split-read analysis, requiring minimum Blat score of 30 on both BSJ sides; (4) **Filtering and validation**: retention of reads mapping to same strand, within 1 Mb distance, non-overlapping by ≥50 bp, and in reverse genomic order; (5) **Annotation and classification**: assignment to RefSeq genes, correction to nearest annotated exons (within 30 bp), and validation against multiple databases (≥95% overlap with annotated circRNAs). The method validates against circBase, circAtlas, and CIRCpedia databases. Further details can be found at [https://github.com/omiics-dk/long_read_circRNA].

Feature	CIRI-long	IsoCirc	circNick-LRS
Input	RCRT-based ONT reads	RCA-based ONT reads	Linearised circRNA ONT reads
Mapping tool	mappy (v2.17)	minimap2 (v2.17)	pblat (v35)
Circular pattern detection	k-mer matching	Tandem Repeat Finder (v 4.0.9)	-
BSJ detection	SPOA, mappy (v2.17), bwapy (v0.1.4)	minimap2 (v2.17)	pblat (v35), bedtools (v2.29.2)
Reference databases	circAtlas	circBase, MiOncoCirc	circBase, circAtlas, CIRCpedia

Tab. 1: *Comparison of key features and computational components across three long-read circular RNA sequencing methods.*

Tool for ONT in-silico generation of long-read sequencing data

NanoSim [6]: is a fast and scalable read simulator that captures the technology-specific features of Oxford Nanopore Technologies (ONT) data. The tool analyzes ONT reads from experimental data to model read features such as length, error profiles, and k-mer biases. In its latest version (v3.0), NanoSim supports the simulation of genomic, transcriptomic (cDNA and direct RNA), and metagenomic reads, accommodating features like intron retention events and chimeric reads. The simulation process involves characterizing the input data to learn these features and then generating synthetic reads that mimic the observed characteristics. NanoSim is implemented in Python and utilizes tools such as minimap2 for alignment and HTSeq for efficient reading of SAM alignment files. Pre-trained models for organisms like *E. coli* and *S. cerevisiae* are available, and users can also train NanoSim on their own datasets to tailor the simulation to specific applications. More details and access to the software can be found at [https://github.com/bcgsc/NanoSim].

Wet-lab dataset

Mouse brain dataset:

We selected the CIRI-long protocol and mouse brain dataset from CIRI-long study as our basis to establish a robust foundation for our benchmarking study. The CIRI-long paper has the highest number of citations and provides the most comprehensively described wet-lab conditions, making it an ideal reference point for comparative analysis. Additionally, mouse circRNA databases contain over one million well-curated annotations, providing robust ground truth data essential for reliable feature extraction and validation of detection accuracy across all three methods.

The FASTQ file from the Zhang et al. study [1] was obtained from the National Genomics Data Center (China National Center for Bioinformation) under the accession number CRA003317 [https://ngdc.cncb.ac.cn/gsa/browse/CRA003317]. This dataset, containing 1,760 total sequences spanning 2 Mbp with read lengths ranging from 110 to 4,635 base pairs and a GC content of 48%, with no sequences flagged as poor quality, served as the foundation for our simulation parameters and error modeling.

Ground truth construction

Our framework integrated two complementary circRNA databases selected for their comprehensive coverage, multi-organism support, and provision of mature circRNA sequences:

circAtlas v.3 (Wu et al., 2020) [4]: An integrated resource cataloging over 3 million circRNAs across 33 tissues and 10 vertebrate species. circAtlas 3.0 provides full-length isoform sequences with extensive functional annotations, including conservation profiles, expression patterns, miRNA and RBP binding sites, and coding potential predictions. It integrates both Illumina and Nanopore sequencing data using standardized nomenclature, facilitating cross-database comparisons and evolutionary studies. Its comprehensive tissue and species coverage provides robust reference data for feature extraction and validation.

circBase (Glažar et al., 2014) [5]: A foundational database that merges and unifies publicly available datasets of circular RNAs identified in eukaryotic cells. circBase provides comprehensive access to circRNA data within genomic context, supporting queries by identifier, gene, or genomic position. The database includes validation scripts for identifying known and novel circRNAs from sequencing data, making it an essential resource for circRNA research and candidate validation. Its broad taxonomic coverage and condition-independent curation make it ideal for establishing general circRNA feature baselines.

We selected these databases based on four key criteria: (1) **Multi-organism support** - both databases provide extensive mouse and human circRNA annotations; (2) **Condition-independent curation** - entries represent general circRNA populations rather than condition-specific datasets; (3) **Sequence availability** - both provide mature circRNA sequences in FASTA format enabling direct feature extraction; and (4) **Comprehensive coverage** - high numbers of validated entries (circBase: >140,000; circAtlas: >3 million) ensure robust statistical analysis. Importantly, we used the intersection of these databases to enhance confidence in our reference dataset, as circRNAs supported by both independent databases are less likely to represent study-specific artifacts and more likely to constitute authentic circRNA sequences.

Genome assembly and annotations: We utilized the GRCm38.p4 mouse genome assembly as our reference genome, selected for compatibility with the circNick-LRS pipeline requirements. Reference genome sequences and genomic annotations were obtained from the GENCODE consortium [13], specifically using the GENCODE version M10 mouse gene annotation file. This version provides comprehensive gene models including protein-coding genes, long non-coding RNAs, and pseudogenes, ensuring complete coverage for circRNA classification and validation across all genomic contexts (exonic, intronic, and intergenic regions).

Data simulation

Feature extraction from databases: To understand the composition of circular RNAs, we have analysed circRNAs found at an intersection between circRNA databases. We developed a computational pipeline to extract comprehensive features from circRNA annotations using Python-based bioinformatics tools (pybedtools, pysam, pandas). The analysis integrated genomic coordinates, sequence information, and transcriptomic context from input BED files, reference genomes (FA), and gene annotation (GTF) files. Our methodology extracted key circRNA characteristics, including genomic location (chromosome, start, end, strand), mature RNA length, and splice site details. We classified splice sites based on canonical motifs (GT-AG, GC-AG, AT-AC) and performed intersectional analysis to annotate gene and transcript types.

Feature extraction from nanopore sequences: To understand the effect of wet lab protocol on sequencing data, we developed a Python-based computational pipeline to extract and visualize detailed sequencing characteristics from FASTQ files. The analysis quantified read metrics including length distributions and repeat patterns present in the wet lab data. Data processing and statistical analysis were performed using Python libraries including pandas and NumPy for handling structured datasets and numerical operations, while visualization utilized Python plotting libraries matplotlib and seaborn for generating distribution plots and statistical comparisons. Rolling circle amplification characteristics were analyzed using Tandem Repeats Finder (TRF v4.09) integrated within our Python framework to detect rolling circle periods, estimate copy numbers and identify repeat patterns.

in-Silico circRNA generation: Building upon our database and wet-lab protocol feature extraction pipelines, we developed an approach to simulate circular RNA types (Fig. 2B). Leveraging genomic features extracted from exonic, intronic, and intergenic regions, we generated four circRNA types: exonic circRNAs (ecircRNAs), circular intronic RNAs (ciRNAs), exon-intron circRNAs (EIciRNAs) and intergenic circRNAs. Key generation features included random sequence extraction, length-constrained generation, splice site preference and rolling circle amplification.

To capture the biological complexity of circRNA isoforms, our simulator incorporates alternative splicing patterns that reflect natural circRNA diversity. For eciRNAs, we implemented exon skipping events with 10% probability for each exon, allowing generation of multiple isoforms from the same genomic locus with varying exonic compositions. EIciRNAs exhibit more complex splicing patterns with 15% probability of exon skipping and 70% probability of intron retention for each intron,

reflecting the characteristic exon-intron structure that defines this circRNA class. ciRNAs are generated from intronic sequences with predominantly single-exon structures (59.4%) representing complete intron circularization, while multi-exonic ciRNAs result from combining multiple intronic regions within the same gene. Intergenic circRNAs are generated from non-genic genomic regions with random positioning and exhibit variable exon counts (35% single-exon, 65% multi-exonic) based on the extracted distribution patterns from our database analysis. These type-specific structural probabilities were derived from our database analysis and ensure that simulated circRNAs exhibit the structural diversity observed in real biological systems, including the presence of multiple isoforms per locus that challenge detection tools differently.

We implemented sequence quality filtering to exclude sequences with more than 10% unidentified nucleotides (N bases or gap in the assembly). The generation process dynamically selected sequence start positions, controlled rolling circle replication defined by user parameters, and captured detailed metadata including transcript identifiers, gene coordinates, and strand information. To validate the simulated circRNAs, we employed BLAT [14] for quality control validation of backsplice junctions. While BLAT is less optimal for long-read alignment compared to minimap2, we selected it for this validation step because its output format displays clearly in genomic browsers (e.g. UCSC [15], IGV [16]), enabling straightforward manual validation of back-splice junctions and visual identification of the characteristic "tail-before-head" pattern of circRNA reads during quality control. It is important to note that BLAT served solely as a quality control validation tool for our simulated data and was not used for precise genomic mapping in the benchmarking analysis, where minimap2 was employed for its superior optimization with nanopore data. This validation step ensures that our simulated circRNAs maintain the structural characteristics necessary for downstream tool evaluation while incorporating the isoform complexity that distinguishes transcriptome-level detection (general boundary identification) from full-length isoform detection (complete sequence reconstruction with accurate internal structure).

in-silico circRNA long-reads data simulation: We utilized NanoSim version 3.1.0 to simulate Nanopore sequencing reads from the circRNA sequences generated in our previous simulation step to use as positive ground truth and from linear RNA sequences to use as negative ground truth (Fig.2C). The simulation uses multiple input files: a control FASTQ file (CRR194180.fq), reference genome (GRCm38.p4), and transcriptome annotation (gencode.vM10). The NanoSim simulation process involved four stages: read analysis for characterizing sequencing properties using the control FASTQ

file with minimap2 alignment, circRNA read simulation generating 200,000 reads from our previously created circRNA.fa file, transcriptome quantification for analyzing expression levels, and linear read simulation generating 200,000 linear RNA reads from transcriptome annotation.

We selected the Guppy basecaller because it represents the most accurate basecaller for ONT data, ensuring our simulated reads reflect contemporary sequencing quality compared to older alternative(Albacore). The cDNA 1D read type was chosen to match the library preparation method used by all benchmarked tools and represents the current ONT standard. We chose minimap2 for alignment during characterization because it provides superior accuracy for long-read transcriptome data compared to alternative LAST aligner. Importantly, we disabled NanoSim's intron retention modeling to prevent conflicts with our custom circRNA-specific simulation scripts, allowing us to control splicing patterns precisely according to our circRNA type-specific generation rather than relying on generic intron retention models. The simulation generated a comprehensive dataset that captured the molecular and sequencing characteristics of circular and linear RNAs, providing a computationally derived representation of Nanopore sequencing data with realistic error profiles and read characteristics.

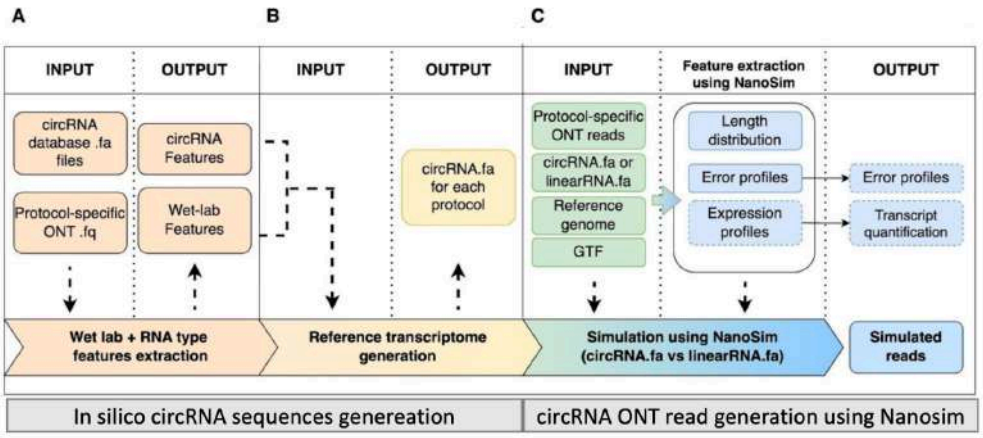


Fig. 2: Computational Workflow for Nanopore Sequencing circRNA Simulation. *The workflow consists of two main phases: circRNA sequence generation (Panels A, B) and nanopore read simulation (Panel C). (A) Feature Extraction: Analysis of circRNA databases and experimental nanopore reads to extract biological parameters including length distributions, splice site patterns, and rolling circle characteristics. (B) CircRNA Generation: Our custom simulator creates biologically realistic circRNA sequences based on extracted features, generating FASTA files with proper circRNA types (eciRNA, EIciRNA, ciRNA, intergenic) and rolling circle structures. (C) NanoSim Read Simulation:*

Conversion of generated circRNA sequences into realistic Oxford Nanopore Technology (ONT) FASTQ reads using NanoSim, incorporating authentic error profiles, quality scores, and read length distributions from control datasets.

Benchmark Dataset

Our generated *in silico* benchmark dataset incorporates the following quantitative and qualitative properties:

Biological properties:

- Organism: *Mus musculus*
- Simulated circRNA count: 7,503 unique circRNAs across four major types (eciRNA, EIciRNA, ciRNA, intergenic)
- Linear transcript reference: 117,667 unique transcripts (GENCODE vM10)
- Total simulated read output: 400,000 reads

Technical parameters (Modeled by NanoSim):

- Sequencing platform: Oxford Nanopore Technologies (ONT)
- Library preparation: cDNA 1D protocol
- Basecaller: Guppy
- Error characteristics: Derived from authentic ONT mouse brain sequencing data
- Homopolymer modeling: Based on actual ONT sequencing artifacts and base-calling limitations
- Mean sequence quality: Q12

Standardisation of tool output

To ensure comprehensive and standardized evaluation, we converted all tool prediction results to BED12 format, which provides a consistent representation of genomic features including chromosomal location, exon structure, and strand information. This normalization allowed for precise comparative analysis across different circRNA detection tools despite different output formats.

Performance metrics

The bedtools intersect approach utilized three parameters to ensure precise genomic feature comparison. The -f (fraction overlap) parameter controls matching stringency by specifying minimum overlap requirements between predicted and ground truth circRNA annotations. We tested three overlap thresholds: 0.25 (lenient, 25% minimum overlap), 0.5 (moderate, 50% overlap), and 0.75 (stringent, 75% overlap) to capture a comprehensive range of detection scenarios.

Our evaluation framework operates at two distinct levels through the -split and -r parameters. Exon-level evaluation, enabled by the -split option, treats each BED12 block as a separate feature, testing whether tools can accurately reconstruct complete internal circRNA structure including proper exon boundaries and splice junction patterns. This stringent approach requires precise isoform reconstruction with correct internal splicing. Transcriptome-level evaluation, conducted without -split, assesses overlap at the whole transcript level, focusing on back-splice junction detection regardless of internal exon structure. The -r (reciprocal overlap) parameter ensures symmetric overlap requirements, preventing asymmetric matches where only partial features overlap.

We utilized bedtools version v2.31.1 to systematically compare BED12 format annotations of *in silico* ground truth against tool outputs. This dual-level approach distinguishes tools that excel at general circRNA boundary detection from those capable of accurate full-length isoform reconstruction.

Performance assessment employed standard metrics across both evaluation levels. True Positives (TP) represent correctly identified circRNAs meeting reciprocal overlap criteria, False Positives (FP) indicate computational artifacts or misannotations predicted as circRNAs, and False Negatives (FN) comprise undetected circRNAs from the simulated dataset. We calculated Precision (TP / [TP + FP]) to measure specificity in avoiding spurious predictions, Recall (TP / [TP + FN]) to assess comprehensive detection capability, and F1 Score as the harmonic mean providing balanced evaluation of overall detection performance across different structural accuracy requirements.

Results

Feature extraction from intersected databases and real data for the generation of realistic simulated datasets results

Our circRNA simulation pipeline integrates Nanopore sequencing data from Zhang et al., database annotations, and computational modeling. (Fig. 3).

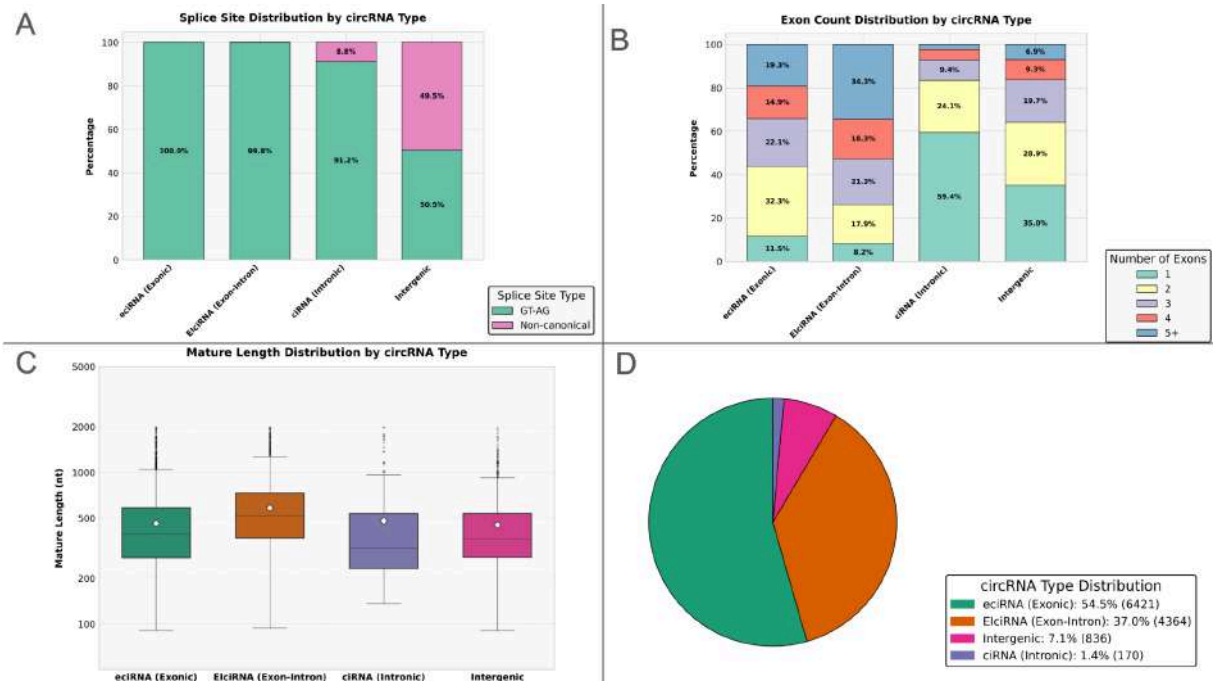


Fig. 3: ***Characterization of Circular RNA molecular features across circRNA types from CircAtlas and CircBase databases****. Panels A-D depict: (A) Splice site composition across circRNA types, (B) Exon count distribution, (C) mature length (exonic concatenation) distribution, and (D) circRNA type distribution.*

Among 11,791 circRNAs common to the two databases, canonical GT-AG splice sites dominated the junction boundaries of exonic, exon-intronic, and intronic circRNAs (100%, 99.8%, and 91.2%, respectively), while intergenic circRNAs showed significantly higher usage of non-canonical splice sites (49.5%) (Fig. 3A).

The exon composition analysis demonstrated substantial variation in complexity across circRNA types (Fig. 3B). Notably, intronic circRNAs (ciRNAs) were predominantly single-exon structures (59.4%), whereas exonic circRNAs (ecircRNAs) showed more diverse exon count distributions with significant proportions containing 2 exons (32.3%) and 3 exons (22.1%). Exon-intron circRNAs (EIciRNAs) exhibited the highest complexity, with 34.3% containing 5 or more exons. Intergenic circRNAs showed an intermediate distribution pattern, with substantial representation across various exon count categories.

Mature length distribution analysis revealed that EIciRNAs possessed the highest median length, followed by ecircRNAs, intergenic, and ciRNAs (Fig. 3C). This pattern reflects the fundamental

structural differences between these circRNA subtypes, with EIciRNAs incorporating both exonic and intronic sequences, contributing to their increased overall length.

When examining the overall distribution of circRNA types in the reference databases (Fig. 3D), ecircRNAs constituted the majority (54.5%, n=6421), followed by EIciRNAs (37.0%, n=4364), intergenic circRNAs (7.1%, n=836), and ciRNAs (1.4%, n=170). This distribution highlights the predominance of exon-derived circular RNAs in the current reference datasets and reflects potential biases in detection methodologies favoring exonic circRNA identification.

These different features were incorporated in our simulation framework based on NanoSim to generate circRNA molecules reflecting biological variations. We verified that the simulated data accurately reflected the characteristics of the input parameters. Specifically, the distributions of circRNA types, splicing patterns, and repeat number variability observed in real Nanopore data and database annotations were preserved in the simulated output

<u>Descriptive comparison of circRNA outputs across tools</u>

We performed a comparison of three circRNA detection tools (CIRI-long, isoCirc, and circNick-lrs) run on a simulated dataset ($n_{circRNAs}$ = 7503, n_{reads} = 400000), to evaluate their performance characteristics and detection biases (Fig. 4).

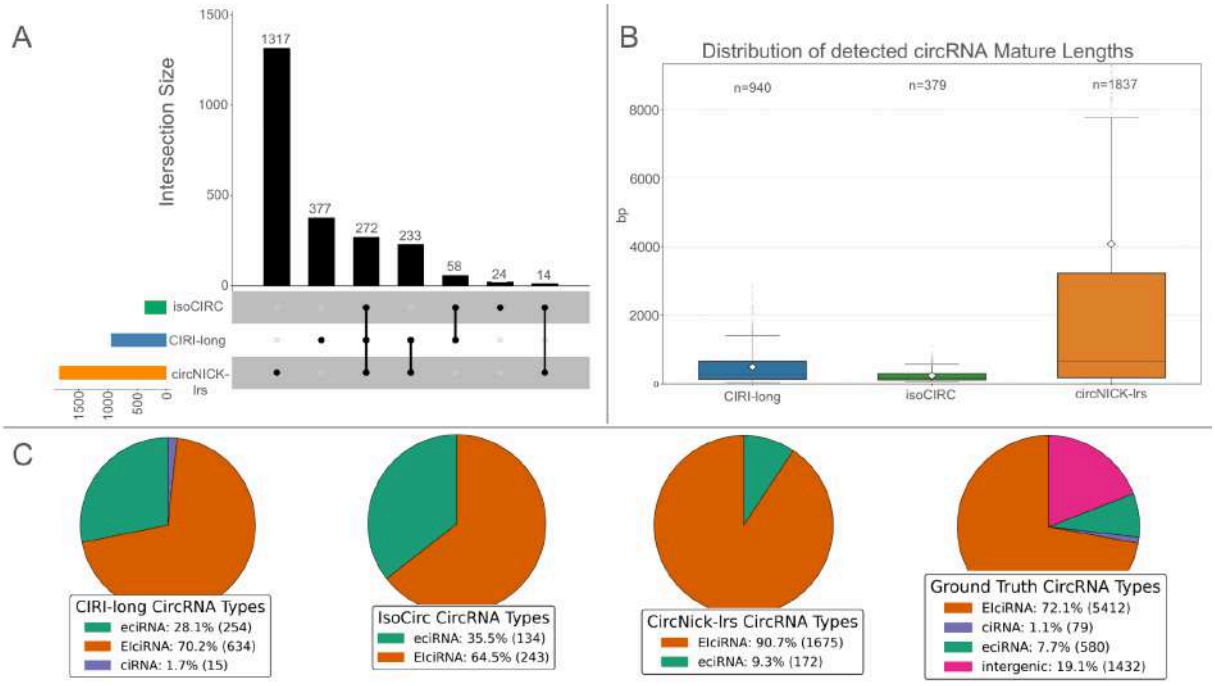


Fig. 4: ***Descriptive comparison of circRNA outputs across tools.*** *Panels A-C depict: (A) intersection analysis showing shared and unique isoform detection across tools, (B) boxplots of circRNAs mature length distribution found by each tool, and (C) proportional distribution of circRNA subtypes identified by each tool compared to ground truth*.

The intersection analysis revealed significant differences in detection capability among the three tools (Fig. 4A). Notably, circNick-Irs demonstrated the highest unique detection capability, identifying 1317 circRNAs not found by other methods. CIRI-long uniquely detected 377 circRNAs, while isoCirc showed the lowest unique detection with only 24 circRNAs. The consensus between all three tools was limited to 272 circRNAs (11.85% of total detections), highlighting the complementary nature of these detection approaches. Additional intersection patterns included 233 circRNAs detected by both circNick-Irs and CIRI-long but missed by isoCirc, and 58 circRNAs identified by both CIRI-long and isoCirc but not by circNick-Irs.

Analysis of mature length distribution revealed differences between the tools (Fig. 4B). CircNick-Irs demonstrated substantially greater capability in detecting longer circRNAs, with a much wider interquartile range extending beyond 3000 bp, and a significant proportion of detections in the higher ranges (25.97% between 1-5 kb, 8.44% between 5-10 kb, and 9.04% above 10 kb). In contrast, both CIRI-long and isoCirc predominantly detected shorter circRNAs. CIRI-long showed an intermediate range capability with 66.60% of detections under 500 bp and 14.47% in the 1-5 kb range, but none above 5 kb. IsoCirc exhibited the most restricted length distribution, with 89.71% of detections under 500 bp, nearly no representation (0.26%) in the 1-5 kb range, and none above 5 kb. This severe limitation in isoCirc can be attributed to its built-in length cutoff at 4000 nucleotides, making it incapable of detecting longer circRNAs without modifying the original code in the TRF (Tandem Repeat Finder) section.

When examining the proportional distribution of circRNA types (Fig. 4C), we observed substantial variation in type bias across tools compared to ground truth. Ground truth composition revealed a distribution of 72.1% EIciRNAs, 7.7% ecircRNAs, 1.1% ciRNAs, and 19.1% intergenic circRNAs. However, all three tools showed significant detection biases. CircNick-Irs demonstrated extreme bias toward EIciRNAs (90.7%) with limited ecircRNA detection (9.3%). CIRI-long showed better balance between EIciRNAs (70.2%) and ecircRNAs (28.1%), with minimal ciRNA detection (1.7%). IsoCirc exhibited a more balanced detection of EIciRNAs (64.5%) and ecircRNAs (35.5%) but completely missed ciRNAs and intergenic circRNAs. Notably, intergenic circRNAs were missed by all three tools despite representing 19.1% of the ground truth, while ciRNAs were detected only by CIRI-long,

suggesting limitations in current circRNA detection methodologies. These findings highlight the importance of tool selection based on the specific circRNA types of interest and suggest that combining multiple tools may provide the most comprehensive circRNA detection.

Performance evaluation of circRNA detection Tools

To evaluate the three circRNA detection tools, we assessed their performance across multiple dimensions including detection accuracy, expression profiling, computational efficiency, and resource requirements (Fig. 5).

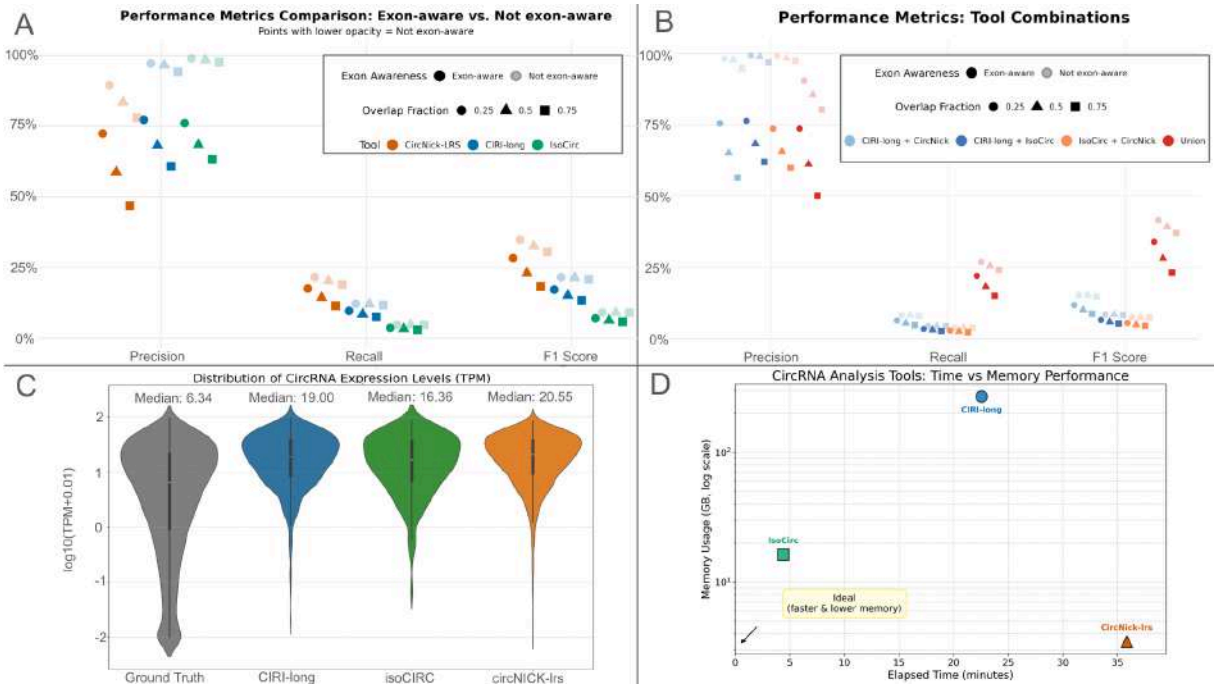


Fig. 5: Performance Evaluation of circRNA Detection Tools on Simulated Datasets. *Panels A-D depict: (A) quantitative comparison of performance metrics, (B) quantitative comparison of performance metrics on combinations and union of tools, (C) violin plot distribution of circRNA expression levels (TPM) detected by each tool compared to ground truth, (D) dot plot showing peak memory usage requirements for each computational approach and displaying processing time per read when analyzing a standardized dataset of 400,000 reads.*

Performance metrics analysis revealed varying detection capabilities across different tools and overlap fractions (Fig. 5A). At the 0.25 fraction, IsoCirc demonstrated the highest precision (99%), followed by CIRI-long (0.97) and CircNick-LRS (0.89). Notably, all tools exhibited low recall, with

CircNick-LRS performing significantly better (0.22) compared to IsoCirc (0.05) and CIRI-long (0.12). As the overlap fraction increased to 0.5, a consistent decline in precision was observed: IsoCirc dropped to 0.98, CIRI-long to 0.97, and CircNick-LRS to 0.83. Recall remained persistently low across all tools: CircNick-LRS at 0.20, CIRI-long at 0.12, and IsoCirc at 0.05. At the 0.75 fraction, the precision decline continued: IsoCirc decreased to 0.98, CIRI-long to 0.94, and CircNick-LRS further dropped to 0.78, with recall values remaining consistently low across all tools: CircNick-LRS at 0.19, CIRI-long at 0.12, and IsoCirc at 0.05. Based on this, we can see that each tool exhibits unique performance profiles. CircNick-LRS shows the most balanced approach with relatively higher recall, CIRI-long maintains consistent performance, while IsoCirc prioritizes extreme precision at the cost of sensitivity.

Performance metrics analysis revealed subtle changes in detection capabilities when implementing the -split option (see M&M for more details) across different tools and overlap fractions (Fig. 5A). The observed performance decline with the -split option reflects the critical challenge of precisely mapping circRNA isoform exon structures. This stringent criterion dramatically reduces detection sensitivity by enforcing a more rigorous matching of exon structures, essentially demanding that the computational tools precisely align with the ground truth of circRNA splicing events. The substantial drops in precision and recall across all tools—most notably for CircNick-LRS—highlight the fundamental difficulty in computationally reconstructing exact circRNA isoform boundaries.

When we merged the circRNA sets of the three tools, significant performance variations emerged (Fig. 5B). The union of all three tools produced the highest recall values across all fractions, substantially outperforming any individual tool or pairwise combination. However, this comprehensive approach came at the cost of reduced precision (0.90) compared to any individual tool or two-tool combination. This trade-off highlights the complementary nature of these detection methods, where combining all tools captures more true positives but introduces additional false positives.

The expression level analysis (Fig. 5C) revealed that all three tools demonstrated significant bias toward detecting highly expressed circRNAs. While the ground truth dataset had a median expression of 6.34 TPM, the detected circRNAs showed substantially higher median expression levels: CircNick-Irs (20.55 TPM), CIRI-long (19.0 TPM), and IsoCirc (16.36 TPM). This observation aligns with detection capabilities, as highly expressed circRNAs generate more supporting reads, making them easier to identify. Minimum expression values were similar across tools, with CIRI-long, CircNick-Irs, and the ground truth dataset showing a minimum expression of 1.0×10^{-2} TPM, while IsoCirc had a slightly higher minimum TPM of approximately 6.0×10^{-2} TPM.

Resource utilization varied dramatically between tools (Fig. 5D), reflecting genuine architectural differences observed under identical testing conditions with 8 threads and identical input datasets. CIRI-long exhibited exceptionally high memory consumption (269.95 GB), representing a substantial computational burden that likely stems from its approach to 1) simultaneous processing of multiple consensus sequences per thread, 2) maintenance of large genome index structures for detecting repetitive patterns, and 3) rolling circle detection algorithms that require extensive k-mer matching operations across all available threads. In contrast, isoCirc (16.28 GB) demonstrated memory-efficient design characteristics, while circNick-LRS (3.46 GB) showed the most efficient memory utilization. Importantly, circNick-LRS's low memory footprint is partly attributable to its single-threaded architecture, which prevents the memory multiplication that occurs with multi-threaded processing, making it more accessible for researchers with limited computational resources but at the cost of processing speed.

Processing speed analysis also revealed significant performance differences that reflect distinct computational architectures. IsoCirc emerged as the clear performance leader, processing 5,454,545 reads per hour (0.66 milliseconds per read), demonstrating superior computational efficiency through optimized multi-threading. CIRI-long demonstrated intermediate efficiency at 1,061,947 reads per hour (3.39 ms/read), while circNick-LRS required substantially more processing time at 669,145 reads per hour (5.38 ms/read), making it the slowest of the three tools. These metrics provide researchers with practical expectations for processing large-scale datasets and help inform tool selection based on computational resource availability, time constraints, and infrastructure capabilities.

Discussion

Our benchmarking study reveals the current state of circRNA detection tools for long-read Nanopore sequencing, highlighting both the promising capabilities and significant challenges in accurately identifying circular RNAs. The analysis underscores the complementary strengths and limitations of current computational approaches, providing critical insights for researchers navigating circRNA detection. Each tool demonstrated distinct performance characteristics that warrant careful consideration.

IsoCirc emerged as the computational efficiency leader, requiring only 0.66 milliseconds per read, making it an attractive option for researchers with limited computational resources. However, this efficiency came at the cost of significant limitations, including a built-in length cutoff at 4000 nucleotides that severely restricts its ability to detect longer circRNAs.

CIRI-long offered a more balanced approach to circRNA detection, with intermediate performance across metrics. However, it imposed a substantial computational burden, consuming an exceptional 269.95 GB of memory—a significant constraint for many research environments.

CircNick-Irs distinguished itself by demonstrating the most comprehensive circRNA detection, particularly for longer circRNAs. It uniquely identified 1,317 circRNAs not found by other methods, with a notable capability to detect circRNAs across diverse length ranges—25.97% between 1-5 kb, 8.44% between 5-10 kb, and 9.04% above 10 kb. However, the tool came with significant limitations, most notably its restriction to only built-in mouse and human reference genomes (mm10 and hg19) with predefined annotation files, preventing its use with custom or alternative genome annotations. Despite its extensive detection capabilities, the tool also required the most processing time at 5.38 ms/read, presenting an additional performance challenge for researchers with large datasets and struggled the most with predicting correct exon structure.

Notably, the extremely low intersection between tools—with only 272 circRNAs (11.85% of total detections) identified by all three methods—strongly suggests that relying on a single detection tool is suboptimal. This stark divergence in detection capabilities demonstrates the methodological challenges in circRNA identification. Each tool essentially acted as a unique lens, capturing distinct aspects of the circRNA landscape that other tools missed. The tools also exhibited significant bias towards detecting highly expressed circRNAs, with median expression levels substantially higher than the ground truth dataset, further complicating comprehensive circRNA characterization.

Our analysis suggests that the choice of tool combination should be guided by specific research priorities. For maximum sensitivity (recall), the union of all tools provides the highest likelihood of detecting true circRNAs, though with more false positives than other approaches. For highest precision, the combination of CIRI-long and isoCIRC maintains the highest precision values across all settings, making it ideal for applications requiring high confidence in detected circRNAs. For balanced performance, while no combination achieves ideal balance, the union approach offers the best compromise between precision and recall, as reflected in its superior F1 scores compared to any two-tool combination.

The study also revealed notable limitations in current detection methodologies when tools can not rely on database validation or standard circRNA composition. All three tools completely missed intergenic circRNAs, which represented 19.1% of the ground truth dataset. Similarly, ciRNAs were

detected only by CIRI-long, suggesting significant gaps in comprehensive circRNA identification across types.

Performance metrics consistently demonstrated low sensitivity across all tools. This highlights the formidable challenge of accurately identifying exact exon structures in circRNA detection.

Our results show mixed consistency with the original publications, reflecting differences in evaluation approaches. CIRI-long's original paper (Zhang et al., 2021) reported higher performance metrics (F1 score of 0.92 for read-level analysis) in simulation studies, but these focused on relatively simple eciRNA structures generated from inner exons without modeling other circRNA types or complex features like exon skipping or intron retention and analysis was conducted on read-level, not on exon/transcriptome level approach we employed here. Our more complex simulation framework, incorporating four circRNA types with alternative splicing patterns, may explain the performance differences from their original validation. The isoCirc publication (Xin et al., 2021) used real datasets for validation, demonstrating high accuracy for full-length circRNA reconstruction across 12 human tissues and HEK293 cells, with emphasis on detecting alternative splicing events within circRNAs that aligns with our observation of isoCirc's precision capabilities. The circNick-LRS study (Rahimi et al., 2021) focused on characterizing circRNA diversity in human and mouse brain samples, with validation primarily through RT-PCR of selected candidates rather than systematic benchmarking metrics. Importantly, none of the prior studies used the similar simulation framework or standardized evaluation metrics across all three tools, making our simulation environment with known ground truth suitable for revealing performance characteristics and tool overlaps that were not systematically evaluated in the original publications.

Limitations of the current study include the exclusion of the circFL-seq [17] tool from the benchmarking analysis. This tool was initially considered for inclusion but was ultimately omitted due to unresolved dependency issues that prevented its execution on cluster.

Notably, other circRNA detection tools in the study also presented significant installation challenges, lacking container environments and experiencing varying difficulties during cluster deployments. These installation difficulties can pose substantial barriers to researchers without bioinformatics expertise, potentially limiting the broader adoption and comparative evaluation of emerging circRNA detection methodologies. This highlights a broader challenge in bioinformatics tool comparison: the practical difficulties of integrating different computational approaches, especially those with complex dependencies or limited ongoing maintenance. Future iterations of this benchmarking study should

continue efforts to incorporate additional circRNA detection tools, potentially developing more standardized approaches to tool integration, documentation, and user accessibility.

In conclusion, this study provides a comprehensive evaluation of circRNA detection tools, emphasizing the need for continued methodological refinement. Researchers must carefully consider the specific requirements of their studies—computational resources, circRNA length, and detection sensitivity—when selecting circRNA detection tools and are advised, when possible, to use their combination.

By integrating wet-lab data, database annotations, and computational modeling, our framework captures circRNA biogenesis complexity and provides a valuable resource for studying circular RNA function and regulation. Our study comes with a customizable framework that allows researchers to analyse database features given mature circRNA sequences from any database as an input or protocol-specific features given .fastq files corresponding to a protocol and tailor circRNA simulation parameters to their specific needs. Our scripts are freely available at https://gitlab.com/bioinfog/circall/nano-circ.

Acknowledgements

The authors would like to thank the Gene Expression and Oncogenesis team, the Dog Genetics team and Biology in Silico 2.0 group (IGDR CNRS UMR6290) for helpful discussions. We also acknowledge the GenOuest bioinformatics core facility (https://www.genouest.org/) for providing the computing infrastructure. This study received financial support from the Ligue National Contre le Cancer (LNCC) Départements du Grand-Ouest. AR is a recipient of a doctoral fellowship from the French Ministry of Higher Education and Research.

References

1. Zhang J, Hou L, Zuo Z, Ji P, Zhang X, Xue Y, et al. Comprehensive profiling of circular RNAs with nanopore sequencing and CIRI-long. Nat Biotechnol. 2021 Jul;39(7):836–45.

2. Xin R, Gao Y, Gao Y, Wang R, Kadash-Edmondson KE, Liu B, et al. isoCirc catalogs full-length circular RNA isoforms in human transcriptomes. Nat Commun. 2021 Jan 12;12(1):266.

3. Rahimi K, Venø MT, Dupont DM, Kjems J. Nanopore sequencing of brain-derived full-length circRNAs reveals circRNA-specific exon usage, intron retention and microexons. Nat Commun. 2021 Aug 10;12(1):4825.

4. Wu W, Ji P, Zhao F. CircAtlas: an integrated resource of one million highly accurate circular RNAs from 1070 vertebrate transcriptomes. Genome Biol. 2020 Dec;21(1):101.

5. Glažar P, Papavasileiou P, Rajewsky N. circBase: a database for circular RNAs. RNA. 2014 Nov;20(11):1666–70.

6. Hafezqorani S, Yang C, Lo T, Nip KM, Warren RL, Birol I. Trans-NanoSim characterizes and simulates nanopore RNA-sequencing data. GigaScience. 2020 Jun 1;9(6):giaa061.

7. Esteller M. Non-coding RNAs in human disease. Nat Rev Genet. 2011 Dec;12(12):861–74.

8. Anastasiadou E, Jacob LS, Slack FJ. Non-coding RNA networks in cancer. Nat Rev Cancer. 2018 Jan;18(1):5–18.

9. Verduci L, Tarcitano E, Strano S, Yarden Y, Blandino G. CircRNAs: role in human diseases and potential use as biomarkers. Cell Death Dis. 2021 May 11;12(5):468.

10. Wang Y, Zhao Y, Bollas A, Wang Y, Au KF. Nanopore sequencing technology, bioinformatics and applications. Nat Biotechnol. 2021 Nov;39(11):1348–65.

11. Drula R, Braicu C, Neagoe IB. Current advances in circular RNA detection and investigation methods: Are we running in circles? Wiley Interdiscip Rev RNA. 2024;15(3):e1850.

12. Li H, Durbin R. Fast and accurate short read alignment with Burrows–Wheeler transform. Bioinformatics. 2009 Jul 15;25(14):1754–60.

13. Mudge JM, Carbonell-Sala S, Diekhans M, Martinez JG, Hunt T, Jungreis I, et al. GENCODE 2025: reference gene annotation for human and mouse. Nucleic Acids Research. 2025 Jan 6;53(D1):D966–75.

14. Kent WJ. BLAT—The BLAST-Like Alignment Tool. Genome Res. 2002 Apr 1;12(4):656–64.

15. Perez G, Barber GP, Benet-Pages A, Casper J, Clawson H, Diekhans M, et al. The UCSC Genome Browser database: 2025 update. Nucleic Acids Research. 2025 Jan 6;53(D1):D1243–9.

16. Robinson JT, Thorvaldsdóttir H, Winckler W, Guttman M, Lander ES, Getz G, et al. Integrative genomics viewer. Nat Biotechnol. 2011 Jan;29(1):24–6.

17. Liu Z, Tao C, Li S, Du M, Bai Y, Hu X, et al. circFL-seq reveals full-length circular RNAs with rolling circular reverse transcription and nanopore sequencing. eLife. 2021 Oct 14;10:e69457.

Strain-dependency of metabolic pathways within 1,494 genomes of lactic bacteria evidenced with Prolipipe, an in silico screening pipeline

Noé ROBERT[1], Jeanne GOT[1], Pauline HAMON-GIRAUD[1], Hélène FALENTIN[2,3] and Anne SIEGEL[1]

[1] Univ Rennes, Inria, CNRS, IRISA, Rennes, 35000, France

[2] INRAE, UMR STLO, Rennes, 35000, France

[3] Institut Agro, UMR STLO, Rennes, 35000, France

Corresponding author: anne.siegel@irisa.fr

Keywords GSM, bacterial strain dependency, in silico pathway screening, lactic acid bacteria (LAB)

Abstract *Genomes from bacteria of interest to the food industry exhibit significant functional variability, yet evaluating this characteristic remains challenging. As public repositories continue to accumulate more genomes, large-scale assessment of metabolic potential emerges as a promising method to highlight this functional variability. The primary challenge lies in automating a workflow to construct metabolic networks from genomes on a massive scale, with enzyme identification in sequences being a critical bottleneck. Here, we present Prolipipe, a pipeline designed for the large-scale assessment of metabolic potential in bacteria, focusing on specific pathways. Given a large dataset of hundreds to thousands of bacterial genomes with known taxonomy and a list of targeted pathways, Prolipipe identifies gene functions through a comprehensive annotation step using three different tools. Then it builds genome-scale metabolic networks for each genome. These networks are then parsed to document the presence or absence of each reaction across all processed genomes and queried for reactions specific to particular pathways. By doing so, the pipeline evaluates the metabolic potential of each genome to carry out the pathway according to its gene content and highlights the best candidates among the large-scale set of genomes. In this study, Prolipipe was applied to 1,494 genomes of lactic acid bacteria, assessing the completion ratio of 761 pathways. We classified pathways according to their maximum completion rate, revealing that 137 pathways can be operated by at least one strain in our dataset. By mapping the identifiers of these pathways onto the pathway ontology graph of the Metacyc database, we highlighted four functional classes of Metacyc (toxin biosynthesis, degradation of aromatic compounds, lipopolysaccharide synthesis and O-antigen biosynthesis) without any of their pathways entirely completed at least once by the strains in the dataset. We then investigated infraspecific variability, a strong indicator of functional variability, and compared the species in our genome dataset based on their tendency to exhibit infraspecific variability. This analysis revealed species potential for strain-dependency, where phenotypes differ among strains of the same species -a feature observable in Prolipipe outputs.*

Introduction

Bacterial genomes of interest to the food industry exhibit a wide range of functional variability, facilitated by mechanisms such as horizontal gene transfer or genomic islands [1]. However, precisely determining the functional roles of these genomes and identifying which species exhibit variability remain challenging. To address these questions, public repository databases are increasingly consolidating more and more genomes ; the NCBI database [2] features 255,669,865 annotated sequences in GenBank format and 4,152,691,448 sequences from WGS studies as of February 2025. This vast amount of data enables large-scale analyses of metabolic capabilities, including studies aimed at selecting organisms encoding enzymes catalyzing specific reactions of interest. By focusing on the reactions within a given pathway, we can identify organisms capable of addressing specific challenges through their metabolism—either by synthesizing valuable compounds or degrading unwanted metabolites. The latter approach requires linking a specific pathway to its reactions, the enzymes catalyzing them, their genes, and identifiers—information stored in databases such as KEGG [3,4,5] or MetaCyc [6].

Although none of these databases provide direct indicators of strain-level metabolic capacities, the greatest challenges lie in the large-scale analysis of potential capabilities using tools designed to construct genome-scale metabolic networks. Few such tools can process thousands of genomes simultaneously, although numerous solutions exist, such as Bactabolize [7], which relies on ModelSEED [8], CarveMe [9], and AutoKEGGRec [10]. Converting annotated genomes into metabolic data requires extensive computational parallelization, which tools such as Mpwt [11] can achieve. However, the primary obstacle remains enzyme detection within genomes. This computationally intensive step can be performed via annotation, as in the RAVEN Toolbox [12], or through targeted searches like BLAST [13], as implemented in GapSeq [14]. Working with large datasets enables the identification of metabolic specificities but requires strictly standardized processing of raw data to prevent annotation biases or inconsistencies arising from non-homogeneous datasets.

To address these challenges, we present Prolipipe [15], a tool for large-scale metabolic profiling of bacteria, focusing on specific pathways. Its strategy is based on raw genomes as input, which are processed through a robust annotation step, followed by targeted and standardized metabolic network construction. To evaluate its scalability and accuracy, Prolipipe was applied to a dataset of 1,494 bacterial genomes, analyzing 761 MetaCyc pathways. The results revealed infraspecific variability in strains of the same species, demonstrating Prolipipe's capacity to uncover strain-specific metabolic capabilities.

Methods

Creation of a large catalogue of bacterial genomes. A dataset of lactic acid bacteria's genomes (LAB) is built in compliance with the following restrictions : the genomes had to be qualified as presumptively safe (QPS) according to the EFSA agreement [16], display no ability for sporulation, or have any known pathogenic effects on plants. As a result, 1,494 LAB genomes were retrieved from the NCBI FTP server in the fna FASTA format. Taxon and strain files were built to complete Prolipipe's inputs by linking species names and taxonomic identifier to files and strain name and status to files, respectively. The list of considered strains is available at (`https://github.com/NoeRobert1/prolipipe_on_LAB`).

Selection and creation of Prolipipe-compatible pathways files. All pathways from the MetaCyc database [6] were extracted, totalizing 3,489 pathways for which the pathway's common name and reactions, the number of reactions, and the taxonomy affiliated with the pathway were referenced. Pathways containing three or more reactions and classified as "observed in bacteria" within MetaCyc were selected, leading to a final set of 761 pathways for analysis.

Prolipipe, a glance of the workflow. We implemented Prolipipe, a Python package whose steps are summarized in Fig. 1. Given a large number of genomes along with a taxon file, a strain file and a list of metabolic pathways, the first step of Prolipipe consists of **genome annotation** using a homogeneous structural and functional annotation approach. This step relies on three tools that have demonstrated their individual role in gene annotation : Prokka version-1.14.6 [17] using the Prokka database (version 20/02/2023) with the *–compliant* flag ; EggNOG-mapper v2.1.12 [18] relying on the eggNOG 5.0 database [19] and using the *–itype genome* and *–genepred prodigal* options ; Bakta 1.8.2 [20] with its own database (version 20/02/2023). Each of these tools generates output files in GenBank format, resulting in three annotation versions per strain. The second step consists of **generating draft genome-scale metabolic networks (GSM).** To achieve this, GBK files are processed via a parallelized application of the PathoLogic algorithm from Pathway Tools [21] using the Python package Mpwt [11]. The previously generated taxon file is provided to refine the analysis, and the *–patho* flag is used to allow PathoLogic inference and the integration of all reactions from the MetaCyc database. At this stage, draft GSMs (three per genome) contain sets of .dat files, the native format of Pathways Tools. These files are managed through the PADMet toolbox [22], a Python package offering a comprehensive suite of tools for metabolic pathway reconstruction and annotated genome comparison. For this study, we used the *pgdb_to_padmet* command along with the *–extract-gene*, *–no-orphan*, and *–source* options for downstream benchmarking, as well as *–padmetRef* option to retain only reactions associated with gene sequences from the draft GSMs and spontaneous reactions. This command generates a single .padmet file per strain and annotation tool. Finally, padmet files corresponding to the same strain are merged to create a consensus GSM per strain using the *padmet_to_padmet* command.

Prolipipe's GSM post-analysis and metabolic networks database generation. GSMs are then queried to assess the metabolic profile of each strain based on the input pathways. Prolipipe compiles a database from all GSMs described within the Padmet files, enabling further comparisons using the *compare_padmet* command from the PADMet toolbox [22]. All results are aggregated into four files, including one named reactions.tsv which documents the presence or absence of reactions across all processed strains. This file is analyzed by Prolipipe to produce pathway-specific result tables, which are then used to generate heatmaps that highlight strain dependency. Alternatively, Prolipipe can produce output files compatible with a SPARQL-endpoint, facilitating the generation of a queryable database. The pipeline is available on GitHub (`https://github.com/AuReMe/prolipipe`).

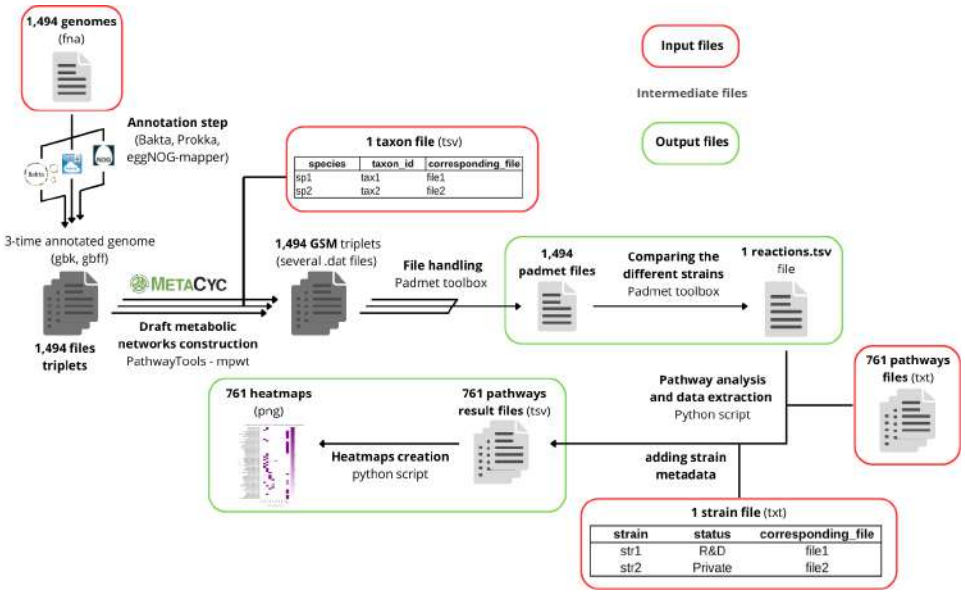


Fig. 1. ***Prolipipe application to our dataset.*** *1,494 genomes along with a taxon file, a strain file and 761 metabolic pathways were processed using Prolipipe, which first annotated the genomes before constructing their genome-scale metabolic networks (GSM). These GSMs were then aggregated into a database and queried to assess the metabolic ability of each strain for each given pathway ; the results were stored in pathway-specific tables which were subsequently used to generate heatmaps.*

Results

Benchmarking of annotation tools used. A catalogue of 1,494 bacterial GSMs has been built using Prolipipe, which records the annotation source when annotating genomes using Eggnog-mapper, Prokka and Bakta. Genomes were grouped by species and the number of reactions ranged from 845 to 1,918 reactions, as shown in Fig. 2. The variance in the number of reactions for each species varied from zero to 13,744, suggesting relative homogeneity within each species group. A bar chart illustrates the contribution of the three annotation tools across all 1,494 GSMs. We observed that the three annotation tools generate consensus predictions accounting for 30 to 43% of the size of the GSMs. Additionally, each tool produced unique predictions that enriched the GSMs.

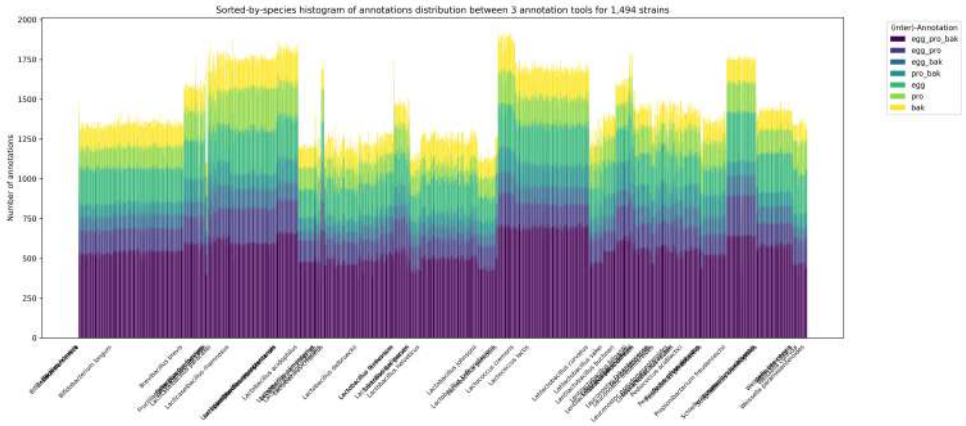


Fig. 2. ***Annotation tools benchmark by evaluating the number of annotations from each tool combination within all 1,494 strains.*** *Species are labelled on their last individual.*

Maximum completion rate per pathway. Prolipipe pipeline was executed on 1,494 bacterial genomes to assess the completion rate of 761 bacterial metabolic pathways for each genome. This rate is calculated as the ratio of reactions within the pathway linked to a gene, as identified by at least one annotation tool, to the total number of reactions in the pathway.
The bar chart in Fig. 3A illustrates the distribution of pathways according to their maximums completion rate, grouped into 10% completion intervals, with a dedicated column for pathways achieving 100% completion. Since placement within the intervals depends heavily on the number of reactions in each pathway, bars are color-coded accordingly. A total of 187 pathways exhibited low completion rates below 10%. This corresponds with the fact that the 1,494 genomes catalogue is composed of lactic acid bacteria which are known for their specialized metabolic properties and inability to perform all known metabolic pathways. Conversely, 137 pathways achieved full completion in at least one strain, as indicated by the [100%] interval column. Furthermore, 29 pathways exhibited a completeness between 80% and 100%, prompting further analysis. Among them, 10 pathways contained at least one lacking an affiliated EC identifier.
Focusing on the 137 pathways completed by at least one strain, the sunburst diagram from Fig. 3B illustrates their distribution within the Metacyc database ontology for all 761 bacterial pathways. Ontology classes shown in color indicate the presence of at least one fully completed within that class, such as galactose degradation (MetaCyc ID : PWY-6317). This pathway aligns with the expected metabolic profile of lactic acid bacteria LAB - known for being able to catabolize this sugar present in milk - used in this study [23]. Prolipipe's table output reveals that 1,384 strains out of 1,494 contain all reactions necessary for this pathway. In contrast, ontology classes depicted in gray indicate categories with no fully completed pathways. These include toxin biosynthesis, degradation of aromatic compounds, lipopolysaccharide synthesis and O-antigen biosynthesis.

Infraspecific variability assessment. Given a pathway, differences in completion rates between strains of the same species indicate the potential for infraspecific metabolic variability. These differences reflect variations in genes annotation related to a specific pathway among strains of the same species. The bar chart depicted in Fig. 4 shows the percentage of infraspecific variability observed over 761 pathways per species (45 species represented by 1,485 strains of the catalogue have more than 1 individual, out of 54 species). This percentage ranges from 0 to 43.9% of all pathways with on one hand *Brevibacillus brevis*'s B showing 334 pathways out of 761, suggesting intra-species metabolic variability. On the other hand *Bifidobacterium longum* shows less variability with a ratio of 25.8% (196 out of 761) while being the most represented species with 145 individuals. This example illustrates that, even though species are not equally represented in the genome catalogue, differences in intra-species metabolic variability potential can still be detected. Infraspecific variability may occur when only a subset of a species' strains complete a given pathway, leading to phenotypic differences between these strains. This phenomenon, known as strain dependency, is discussed further below.

Focusing on L-arginine biosynthesis through acetyl cycle. The heatmap in Fig. 5 depicts the completion rate of the L-arginine biosynthesis through the acetyl cycle, a metabolic pathway enabling organisms to produce L-arginine from L-glutamine and L-glutamate through 9 different reactions. Completion ratios were obtained for the 1,494 strains (covering 54 species) of our dataset using Prolipipe. The X-axis divides completion percentage values into 10%-intervals, strains are ordered

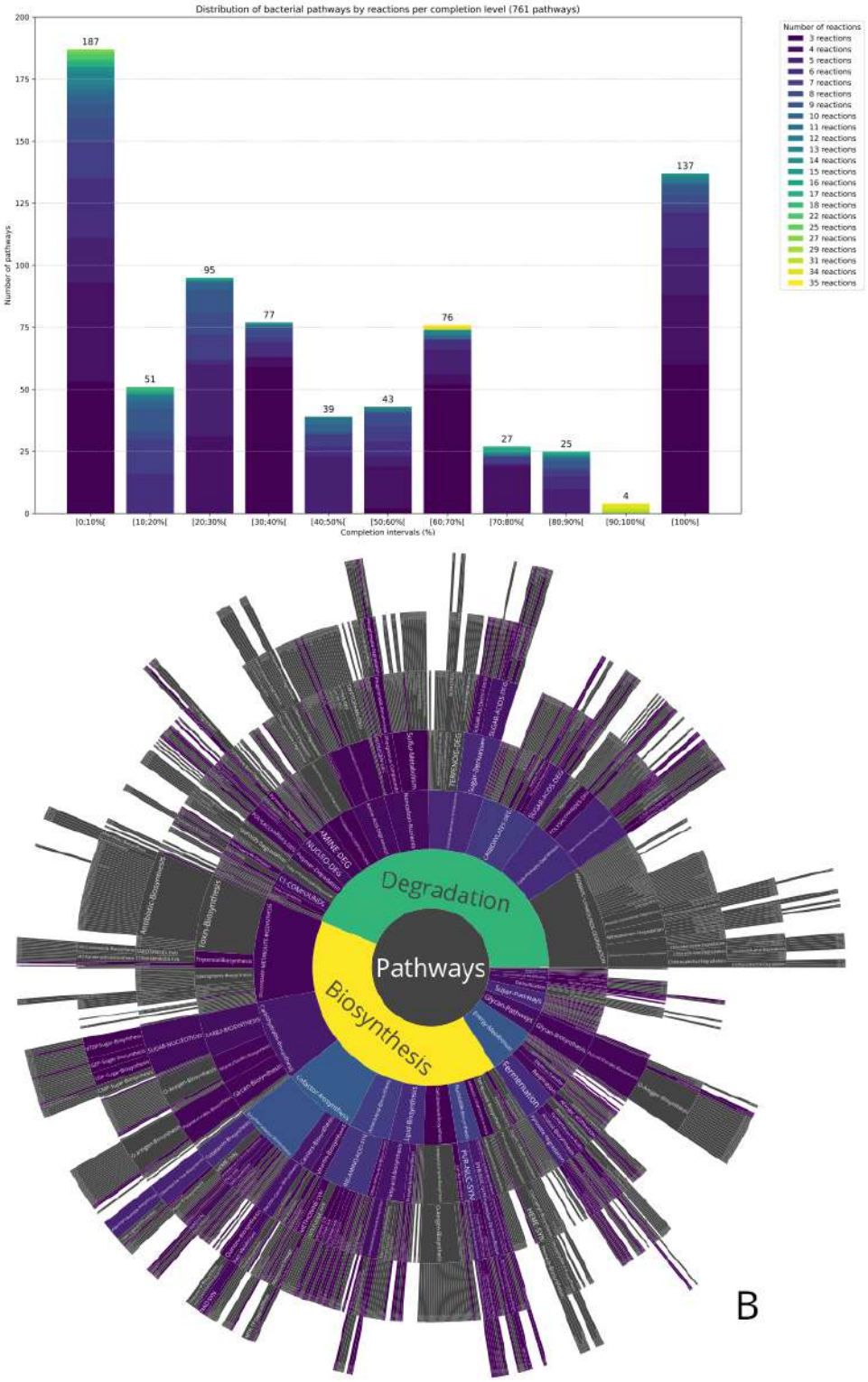


Fig. 3. ***A) Bar chart of pathway completion rate distribution.*** *761 pathways are dealt depending on their completion rate which is splitted into 10%-completion brackets, with an eleventh column for strictly complete pathways. The number of reactions per pathway is indicated with the shade of color.*
B) Sunburst diagram on pathways ontology. *137 pathways are found completed by at least one of the dataset's strain and their repartition in the MetaCyc ontology is screened on all the 761 processed pathways, with colored areas having at least one of these completed pathways.*

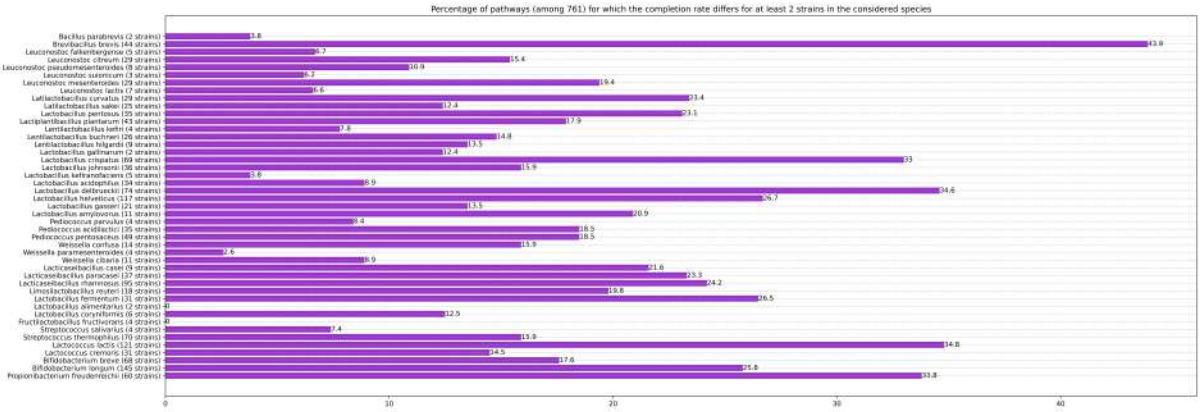


Fig. 4. ***Bar-chart of infraspecific variability frequency.*** *Each species represented by more than one individual is represented by a bar, with the horizontal axis indicating the ratio of observed infraspecific variability across all 761 analyzed pathways. Infraspecific variability is identified when two strains within a species exhibit completion rates for a given pathway that differ by at least 10%. This suggests that the strains in question have different enzymatic capabilities associated with the metabolic pathway.*

taxonomically and grouped into the 54 species of the dataset, and color shade represents the proportion of strains within a species whose completion rate falls within a specific interval.

A total of 23 species exhibited a fully colored cell in the 100% completion column, indicating that all their representative strains achieved full pathway completion. These species are strong candidates for carrying out the pathway and producing L-arginine under suitable environmental conditions. For example, *Streptococcus thermophilus* (70 strains) achieved 100% completion, confirming its status of candidate for carrying out the pathway entirely.

In contrast, 28 species did not contain any strain with 100% completion rate. Among these, some exhibited high variability in completion rates. *Lactobacillus delbrueckii*, for instance, displayed five different completion rates represented by several cells of different shades on a same line : two strains have genomes annotated with the genes of one reaction ; 21 strains are associated with two reactions ; nine strains are associated with three reactions ; six strains are associated with four reactions and 36 strains are associated with five reactions. For these 28 species, the ability to synthesize L-arginine through the acetyl cycle is considered unlikely.

Finally, three species -*Leuconostoc lactis*, *Lactobacillus fermentum* and *Lactococcus cremoris*- displayed infraspecific variability. In *Leuconostoc lactis*, 6 out of 7 strains (85.7%) achieved 100% completion. Similarly, in *Lactobacillus fermentum* and *Lactococcus cremoris*, 87.1% and 90% of strains, respectively, reached full completion. These findings indicate that most strains within these species have the genetic potential to synthesize L-arginine through the acetyl cycle. However, other strains associated with these species have low completion rates. This suggests that the considered species are candidate to be strain-dependent with respect to the phenotype of arginine production [24].

Discussion

From the previous figures, we demonstrated that Prolipipe has significant potential for studying large-scale datasets of bacterial genomes and enables two key analyses. The first key analysis is a large-scale detection of pathways which are promising candidates for the effective production or degradation of metabolites based on their genome content. In this context, Prolipipe can be used for

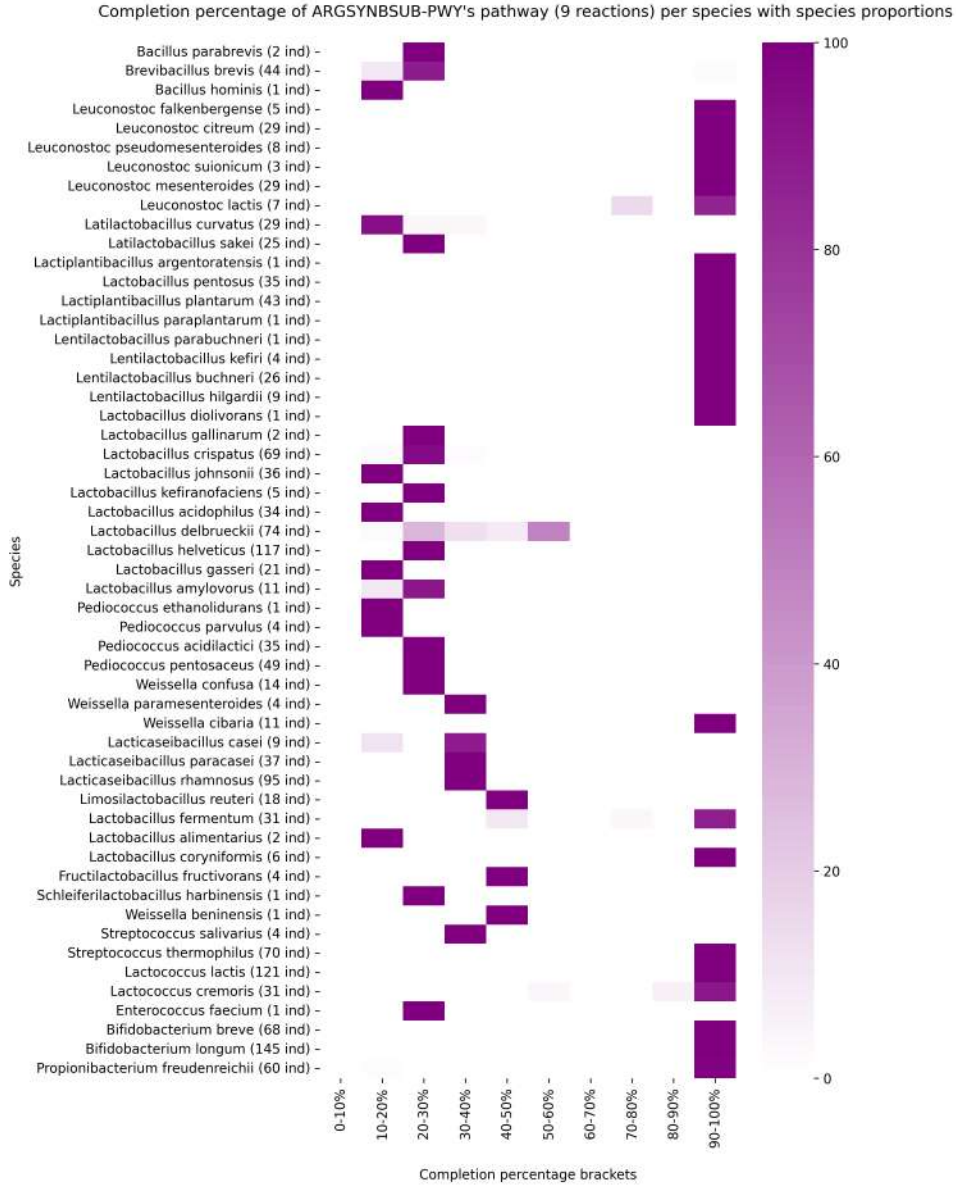


Fig. 5. ***Completion heatmap of ARGSYNBSUB-PWY (Metacyc ID of L-arginine biosynthesis pathway through acetyl cycle) within all 1,494 genomes.*** *Completion ratio among the 54 species (represented in line, headcount given in label) is dealt on 10%-completion brackets, proportion of individuals inside a species in shades of purple.*

genome mining by identifying bacteria with the highest potential for pathways of interest, such as biosynthesis of target compounds or degradation pathways. This process is facilitated by the use of sunburst diagrams that utilize pathway ontology to identify over- and under-represented functional classes. An example was provided with galactose degradation, where Prolipipe output indicated that this pathway is predominantly present in our strains' genomes, which corroborates the literature [23] and demonstrates its compatibility with genome mining approaches.

A second key analysis is the detection of infraspecific variability in pathway completion rate within large-scale genome catalogues. This analysis can be extended to comparison between species, provided that the dataset is homogeneous in terms of species representation. Indeed, in our current catalogue, species headcounts range from 1 individual to 145, thereby complicating the discrimination between true infraspecific variability, known to exist within LAB [25] and biases arising from unequal species representation. Nevertheless, these two aspects -the detection of accessible pathways and infraspecific variability- are directly visualized in readily available, pathway-specific heatmaps which serve as valuable tools for more detailed studies of metabolic profiles related a given metabolic pathway within a bacterial dataset. Moreover, such displays can reveal candidates for strain-dependency, defined as differences in pathway completion rate among strains of the same species, with a subset of the specie's strains achieving 100% completion of the pathway. It is important to note that, since this analysis is based solely on genomic data, the associated phenotypic traits must be validated experimentally.

The investigation of nearly completed pathways revealed the risk of false negatives due to gaps in the MetaCyc database, as some reactions lack an Enzyme Commission (EC) number to link annotation to metabolic data. This limitation persists despite the explicit inclusion of spontaneous reactions during the construction of metabolic networks. Additionally, there is a risk of false positives arising from the triple annotation process, where pseudogenes may be erroneously identified as functional genes. Such risk can be mitigated by adjusting the annotation tools' parameters and increasing the stringency of gene coverage criteria during gene detection. However, due to these potential errors, Prolipipe should be considered as a preliminary tool for assessing diversity potential. Its primary role is to filter out the least promising strains, without guaranteeing quality or exhaustiveness -questions that can be addressed once the genomes catalogue is significantly reduced.

Prolipipe's capabilities are currently being extended to metagenomics, where metabolic capability assessment would not be limited to individual organisms but would encompass entire bacterial communities, facilitating research on metabolic complementarity. This extension is expected to be valuable for the development of bacterial consortia to address challenges that single strains cannot overcome. Additionally, Prolipipe is being adapted for eukaryotic analysis, specifically for assessing metabolic machineries in yeasts.

Acknowledgements

Authors thank the bioinformatics core facility Genouest (https://www.genouest.org/).

Data, scripts, code, and supplementary information availability

Scripts and code are available online: https://github.com/AuReMe/prolipipe
Metadata of strains are available online : https://github.com/NoeRobert1/prolipipe_on_LAB

Conflict of interest disclosure

The authors declare that they comply with the PCI rule of having no financial conflicts of interest in relation to the content of the article.

References

[1] Stefanovic E, Fitzgerald G, McAuliffe O. Advances in the genomics and metabolomics of dairy lactobacilli: A review. Food Microbiology. 2017 Feb;61:33-49. Available from: https://www.sciencedirect.com/science/article/pii/S0740002016306013.

[2] Sayers EW, Beck J, Bolton EE, Brister JR, Chan J, Comeau DC, et al. Database resources of the National Center for Biotechnology Information. Nucleic Acids Research. 2024 Jan;52(D1):D33-43.

[3] Kanehisa M, Furumichi M, Sato Y, Matsuura Y, Ishiguro-Watanabe M. KEGG: biological systems database as a model of the real world. Nucleic Acids Research. 2025 Jan;53(D1):D672-7.

[4] Kanehisa M. Toward understanding the origin and evolution of cellular organisms. Protein Science: A Publication of the Protein Society. 2019 Nov;28(11):1947-51.

[5] Kanehisa M, Goto S. KEGG: kyoto encyclopedia of genes and genomes. Nucleic Acids Research. 2000 Jan;28(1):27-30.

[6] Caspi R, Billington R, Keseler IM, Kothari A, Krummenacker M, Midford PE, et al. The MetaCyc database of metabolic pathways and enzymes - a 2019 update. Nucleic Acids Research. 2020 Jan;48(D1):D445-53. Available from: https://doi.org/10.1093/nar/gkz862.

[7] Vezina B, Watts SC, Hawkey J, Cooper HB, Judd LM, Jenney AW, et al. Bactabolize is a tool for high-throughput generation of bacterial strain-specific metabolic models. eLife. 2023 Oct;12:RP87406. Publisher: eLife Sciences Publications, Ltd. Available from: https://doi.org/10.7554/eLife.87406.

[8] Seaver SMD, Liu F, Zhang Q, Jeffryes J, Faria JP, Edirisinghe JN, et al. The ModelSEED Biochemistry Database for the integration of metabolic annotations and the reconstruction, comparison and analysis of metabolic models for plants, fungi and microbes. Nucleic Acids Research. 2021 Jan;49(D1):D575-88.

[9] Machado D, Andrejev S, Tramontano M, Patil KR. Fast automated reconstruction of genome-scale metabolic models for microbial species and communities. Nucleic Acids Research. 2018 Sep;46(15):7542-53. Available from: https://doi.org/10.1093/nar/gky537.

[10] Karlsen E, Schulz C, Almaas E. Automated generation of genome-scale metabolic draft reconstructions based on KEGG. BMC bioinformatics. 2018 Dec;19(1):467.

[11] AuReMe/mpwt. AuReMe; 2025. Original-date: 2018-12-11T09:01:00Z. Available from: https://github.com/AuReMe/mpwt.

[12] Wang H, Marcišauskas S, Sánchez BJ, Domenzain I, Hermansson D, Agren R, et al. RAVEN 2.0: A versatile toolbox for metabolic network reconstruction and a case study on Streptomyces coelicolor. PLOS Computational Biology. 2018 Oct;14(10):e1006541. Publisher: Public Library of Science. Available from: https://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1006541.

[13] Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. Journal of Molecular Biology. 1990 Oct;215(3):403-10.

[14] Zimmermann J, Kaleta C, Waschina S. gapseq: informed prediction of bacterial metabolic pathways and reconstruction of accurate metabolic models. Genome Biology. 2021 Mar;22(1):81. Available from: https://doi.org/10.1186/s13059-021-02295-1.

[15] AuReMe/prolipipe. AuReMe; 2025. Original-date: 2023-04-14T14:49:42Z. Available from: https://github.com/AuReMe/prolipipe.

[16] Update of the list of QPS-recommended biological agents intentionally added to food or feed as notified to EFSA 15: suitability of taxonomic units notified to EFSA until September 2021 | EFSA; 2022. Section: Scientific outputs. Available from: https://www.efsa.europa.eu/en/efsajournal/pub/7045.

[17] Seemann T. Prokka: rapid prokaryotic genome annotation. Bioinformatics (Oxford, England). 2014 Jul;30(14):2068-9.

[18] Cantalapiedra CP, Hernández-Plaza A, Letunic I, Bork P, Huerta-Cepas J. eggNOG-mapper v2: Functional Annotation, Orthology Assignments, and Domain Prediction at the Metagenomic Scale. Molecular Biology and Evolution. 2021 Dec;38(12):5825-9.

[19] Huerta-Cepas J, Szklarczyk D, Heller D, Hernández-Plaza A, Forslund SK, Cook H, et al. eggNOG 5.0: a hierarchical, functionally and phylogenetically annotated orthology resource based on 5090 organisms and 2502 viruses. Nucleic Acids Research. 2019 Jan;47(D1):D309-14.

[20] Schwengers O, Jelonek L, Dieckmann MA, Beyvers S, Blom J, Goesmann A. Bakta: rapid and standardized annotation of bacterial genomes via alignment-free sequence identification. Microbial Genomics. 2021;7(11):000685. Publisher: Microbiology Society,. Available from: https://www.microbiologyresearch.org/content/journal/mgen/10.1099/mgen.0.000685.

[21] Karp PD, Midford PE, Billington R, Kothari A, Krummenacker M, Latendresse M, et al. Pathway Tools version 23.0 update: software for pathway/genome informatics and systems biology. Briefings in Bioinformatics. 2021 Jan;22(1):109-26.

[22] AuReMe/padmet. AuReMe; 2025. Original-date: 2019-01-28T15:12:24Z. Available from: https://github.com/AuReMe/padmet.

[23] Wang Y, Wu J, Lv M, Shao Z, Hungwe M, Wang J, et al. Metabolism Characteristics of Lactic Acid Bacteria and the Expanding Applications in Food Industry. Frontiers in Bioengineering and Biotechnology. 2021 May;9. Publisher: Frontiers. Available from: https://www.frontiersin.org/journals/bioengineering-and-biotechnology/articles/10.3389/fbioe.2021.612285/full.

[24] Bringel F, Hubert JC. Extent of genetic lesions of the arginine and pyrimidine biosynthetic pathways in Lactobacillus plantarum, L. paraplantarum, L. pentosus, and L. casei: prevalence of CO(2)-dependent auxotrophs and characterization of deficient arg genes in L. plantarum. Applied and Environmental Microbiology. 2003 May;69(5):2674-83.

[25] Thierry A, Valence F, Deutsch SM, Even S, Falentin H, Le Loir Y, et al. Strain-to-strain differences within lactic and propionic acid bacteria species strongly impact the properties of cheese–A review. Dairy Science & Technology. 2015 Nov;95(6):895-918. Available from: https://doi.org/10.1007/s13594-015-0267-9.

MethMotif 2024 Suite Reveals the Epigenetic Blueprint of Context-Specific Transcription Factor Binding Sites

Matthew DYER[1], Quy Xiao Xuan LIN[2], Denis THIEFFRY[2,3] and Touati BENOUKRAF[1,2]

1 Division of BioMedical Sciences, Faculty of Medicine, Memorial University of Newfoundland, 300 Prince Philip Drive, NL A1B 3V6, St. John's, Canada

2 Cancer Science Institute of Singapore, National University of Singapore, 14 Medical Drive, 117599, Singapore, Singapore

3 Département de Biologie de l'École Normale Supérieure, PSL Research University, 46 Rue d'Ulm, 75005, Paris, France

Corresponding Author: tbenoukraf@mun.ca

Paper Reference: Dyer *et al.* (2024) MethMotif.Org 2024: a database integrating context-specific transcription factor-binding motifs with DNA methylation patterns. *Nucleic Acids Research*. https://doi.org/10.1093/nar/gkad894

Keywords

DNA Methylation, Transcription Factors, Cofactor Segregation, Epigenetic Regulation

Abstract

MethMotif (https://methmotif.org) is a publicly available database that provides a comprehensive repository of transcription factor (TF)-binding profiles enriched with DNA methylation patterns. Since its inception in 2019, the platform has evolved to incorporate expanded datasets and advanced functionalities, deepening our understanding of context-specific TF functions. In its 2024 release, MethMotif expands its initial collection from 509 to over 700 position weight matrices (PWMs), all annotated with DNA methylation profiles. A key advancement of this update is the segregation of TF-binding motifs based on cofactors and DNA methylation status, allowing researchers to explore how gene ontology (GO) annotations and TF target genes can differ under varying cofactor contexts. MethMotif now supports two additional species: *Mus musculus* and *Arabidopsis thaliana*, broadening its applicability for comparative and translational research. By incorporating cofactor-based binding motifs, methylation profiles, and precomputed GO enrichments, MethMotif stands out as the first and only TF-binding motif

database to integrate context-specific PWMs with epigenetic information, thus enabling deeper insights into the regulatory mechanisms governing gene expression.

Highlight

MethMotif [1,2] bridges a critical gap in epigenomic gene regulation analysis by integrating transcription factor binding profiles with DNA methylation patterns, offering a unified framework for exploring context-specific regulatory mechanisms. For the JOBIM community, this resource offers three principal advantages:

- Context-Specific Integration: MethMotif incorporates TF-DNA interaction data and DNA methylation patterns, highlighting how cofactors and epigenetic states modulate TF binding. This integrative view is vital for understanding dynamic regulatory events in health and disease.
- Enhanced Species Coverage: In addition to Human, the 2024 release includes data for *Mus musculus* and *Arabidopsis thaliana*, in addition to human cell lines, enabling cross-species comparisons of DNA methylation effects on TF binding and regulatory network evolution.
- User-Focused Tools and Batch Querying: New data visualization modules and batch-query functionalities allow users to systematically explore transcription factor binding sites (TFBS) methylation status, cofactor interactions, and gene ontology enrichments. The TFregulomeR R package [3] further extends these capabilities for custom analyses.

By expanding PWMs, integrating cofactor-based motif segregation, and incorporating pioneer TF annotations, MethMotif 2024 remains a unique and powerful platform for dissecting the epigenetic dimensions of transcriptional regulation.

References

1. Xuan Lin QX, Sian S, An O, Thieffry D, Jha S, Benoukraf T. MethMotif: an integrative cell specific database of transcription factor binding motifs coupled with DNA methylation profiles. Nucleic Acids Res. 2019 Jan 8;47(D1):D145–54.
2. Dyer M, Lin QXX, Shapoval S, Thieffry D, Benoukraf T. MethMotif.Org 2024: a database integrating context-specific transcription factor-binding motifs with DNA methylation patterns. Nucleic Acids Res. 2023 Oct 18;gkad894.
3. Lin QXX, Thieffry D, Jha S, Benoukraf T. TFregulomeR reveals transcription factors' context-specific features and functions. Nucleic Acids Res. 2020 Jan 24;48(2):e10.

AntiBody Sequence Database

Simon MALESYS[1], Rachel TORCHET[1], Bertrand SAUNIER[2] and Nicolas MAILLET[1]

[1] Institut Pasteur, Université Paris Cité, Bioinformatics and Biostatistics Hub, F-75015 Paris, France
[2] Institut Pasteur, Université Paris Cité, Unité de Virologie Structurale, CNRS UMR 3569, F-75015 Paris, France

Corresponding author: nicolas.maillet@pasteur.fr

Reference paper: **Malesys *et al.* (2024) AntiBody Sequence Database. *NAR Genomics and Bioinformatics*.** https://doi.org/10.1093/nargab/lqae171

Keywords antibody variable regions, protein sequence, machine learning, database, standardized

Abstract *Antibodies play a crucial role in the humoral immune response against health threats, such as viral infections. Training AI (Artificial Intelligence) models, for example to assist in developing sero-diagnostics or antibody-based therapies, requires building datasets according to strict criteria, to include as many standardized antibody sequences as possible. However, the available sequences are scattered across partially redundant databases and compiling them into a single non-redundant standardized dataset has hitherto remained a challenge.*

Here, we present ***ABSD*** *(AntiBody Sequence Database, https://absd.pasteur.cloud) which contains data from major publicly-available resources (abYbank, CATNAP-HIV, CoV-AbDab, GeneBank, IMGT, KABAT, OAS, PDB, PLAbDab, PairedNGS, SACS, SAbDab, UniProt...), creating the largest* ***standardized***, ***automatically updated*** *and* ***non-redundant*** *(i.e., each antibody sequence stored in the database is unique) source of public antibody sequences for different species.*

While ABSD contains over ***1,350,000*** *antibody sequences today, trillions of them may circulate in the human population. This limitation is unlikely to be resolved anytime soon, but diversity might matter more than sheer number. In the article, we demonstrate that, at least regarding IGHV regions, our methodology does not seem to have introduced a strong bias in the selection of antibody sequences towards specific gene clusters, compared to a classic human repertoire.*

When training deep learning models, the ***uniqueness and representativeness of the input data*** *is likely essential for most applications. In this regard, ABSD will help mirror the human repertoire by providing,* ***as broadly as possible and without bias****, unique antibody sequences with realistic proportions.*

Finally, ABSD is a dynamic and adaptive database, designed for automatic updates and easy upgrades. This user-friendly and open website enables users to generate lists of antibodies based on selected criteria and download the unique sequence pairs of their variable regions.

Highlight

For AI training and validation in biology, quality of the data likely matters more than quantity. However, the required standards are not always met, as for the plethora of data accumulating in databases, the accuracy and the relationship with biological dimensions are not always ascertained, probably rendering the accessed information not always as pertinent as it should.

To gain insight into the diversity of antibody sequences mobilized in response to specific biological processes or diseases, we wanted to train AI models. We therefore asked ourselves: what is required

to properly train these models? We identify the need for non-redundant and standardized sequence data with clean annotation. Furthermore, it was essential to have trust in these data (acquisition methods, origin of sample, species, 3D structure, etc.), hence, to keep a record of the origins of each sequence. The goal was not necessarily to accumulate as much sequences as possible —the amount of data accumulated so far being anyway very limited compared to reality— but rather to organize a vast amount of data to extract a subset that retains the original diversity, complexity and biological information.

The pertinence of a data source is as important as the quality of biological information it contains. Starting with the merge of high-quality databases (e.g., IMGT and the PDB), we gradually included more databases while refining potential use-cases and addressing the challenges posed by new additions. Ultimately, we developed a fully automated pipeline where adding new databases is straightforward, regardless of their original data quality. This pipeline ensures that each antibody in ABSD is unique, standardized, annotated, and has a direct link to each database from which it was sourced.

We believe ABSD is of interest to the JOBIM community because : 1/ it is a new resource that is directly useful for the bioinformatics and immunology communities, 2/ it introduces a large dataset that, because of its quality and scalable representativeness, is likely well-suited for AI training, and 3/ it demonstrates the challenges, engineering, and methodologies used to achieve a homogeneous and coherent merge of heterogeneous data from numerous public databases, potentially applicable to other bioinformatics domains.

The Pfam protein families database: embracing AI/ML

Typhaine PAYSAN-LAFOSSE[1], Antonina ANDREEVA[1], Matthias BLUM[1], Sara Rocio CHUGURANSKY[1], Tiago GREGO[1], Beatriz LAZARO PINTO[1], Gustavo A SALAZAR[1], Maxwell L BILESCHI[2], Felipe LLINARES-LÓPEZ[3], Laetitia MENG-PAPAXANTHOS[3], Lucy J COLWELL[2], Nick V GRISHIN[4], R Dustin SCHAEFFER[4], Damiano CLEMENTEL[5], Silvio C E TOSATTO[5], Erik SONNHAMMER[6], Valerie WOOD[7], Alex BATEMAN[1]

1 European Molecular Biology Laboratory, European Bioinformatics Institute (EMBL-EBI), Wellcome Genome Campus, Hinxton CB10 1SD, UK

2 Google DeepMind, 355 Main Street, Cambridge, MA 02142, USA

3 Google DeepMind, Brandschenkestr. 110, Zurich 8002, Switzerland

4 Department of Biophysics, University of Texas Southwestern Medical Center, 5323 Harry Hines Blvd., Dallas, TX75390, USA

5 Department of Biomedical Sciences, University of Padova, Via 8 Febbraio, 2, 35122 Padova, Italy

6 Department of Biochemistry and Biophysics, Stockholm University, Science for Life Laboratory, Tomtebodavägen 23A, 17165 Solna, Sweden

7 Department of Biochemistry, University of Cambridge, Hopkins Building Downing Site, Tennis Court Road, Cambridge CB2 1QW, UK

Corresponding Author: typhaine@ebi.ac.uk

Paper Reference: Paysan-Lafosse *et al.* (2024) The Pfam protein families database: embracing AI/ML. *Nucleic Acids Research.* https://doi.org/10.1093/nar/gkae997

Keywords

Protein family, Sequence alignment, Deep learning

Abstract

The Pfam protein families database is a comprehensive collection of protein domains and families used for genome annotation and protein structure and function analysis (https://www.ebi.ac.uk/interpro/). This update describes major developments in Pfam since 2020, including decommissioning the Pfam website and integration with InterPro, harmonization with the ECOD structural classification, and expanded curation of metagenomic, microprotein and repeat-containing families. We highlight how AlphaFold structure predictions are being leveraged to refine domain boundaries and identify new domains. New families discovered

through large-scale sequence similarity analysis of AlphaFold models are described. We also detail the development of Pfam-N, which uses deep learning to expand family coverage, achieving an 8.8% increase in UniProtKB coverage compared to standard Pfam. We discuss plans for more frequent Pfam releases integrated with InterPro and the potential for artificial intelligence to further assist curation. Despite recent advances, many protein families remain to be classified, and Pfam continues working toward comprehensive coverage of the protein universe.

Highlight

The recent enhancements to Pfam, an invaluable, open-source, and widely recognised resource in the scientific community, offer significant benefits for the JOBIM community. For decades, Pfam has served as a cornerstone for protein research, earning widespread trust across disciplines. The integration with InterPro and alignment with ECOD structural classification further strengthens this essential resource for structural and functional genomics investigations. In particular, the application of AlphaFold structure predictions to refine domain boundaries and identify novel domains demonstrates how cutting-edge AI technologies can advance protein annotation. The development of Pfam-N, which achieves an impressive 8.8% increase in UniProtKB coverage through deep learning approaches, represents a meaningful advancement in our understanding of the protein universe. These AI-driven approaches exemplify the type of interdisciplinary integration that the JOBIM community values. Additionally, Pfam's expanded curation of metagenomic and microprotein families addresses key emerging areas of interest. These developments provide the JOBIM community with enhanced tools for genome annotation and protein characterisation, facilitating discoveries across the molecular life sciences.

Session 7: Workflows, Reproducibility, and Open Science

Assessing bioinformatics software annotations: bio.tools case-study

Ulysse LE CLANCHE[1], Sarah COHEN BOULAKIA[2], Yann LE CUNFF[1], Olivier DAMERON[1] and Alban GAIGNARD[3,4]

[1] Université Rennes, Inria, CNRS, IRISA—UMR 6074, Rennes 35000, France
[2] Université Paris-Saclay, CNRS, Laboratoire Interdisciplinaire des Sciences du Numérique, 91405, Orsay, France
[3] Nantes Université, CNRS, INSERM, l'institut du thorax, F-44000 Nantes, France
[4] IFB-core, Institut Français de Bioinformatique (IFB), CNRS, INSERM, INRAE, CEA, 91057 Evry, France

Corresponding author: ulysse.le-clanche@irisa.fr

Keywords Ontologies, Semantic annotations, FAIR bioinformatics software, EDAM, bio.tools

Abstract *Reproducibility and reuse of digital bioinformatics resources are essential for the development of open and cumulative science, in line with FAIR principles. To search and reuse bioinformatics tools, scientists need to be confident enough with the reliability of their annotations. Our study focuses on the quantitative and qualitative evaluation of semantic annotations in the bio.tools registry, which serves more than 30,000 bioinformatics tool descriptions, annotated with the EDAM ontology. In this work we propose to study how the EDAM ontology is used to categorize software based on scientific disciplines and the kind of data processing they allow. We also evaluate how qualitative are the annotations based on Shannon entropy. We emphasize that a particular attention should be given to the whole set of inherited annotations, from the used ontology. Our results underline the need for automatic tools to support annotation curation, reducing the annotation cost for domain experts. This study is a preliminary work aimed designing novel annotation approaches based on the combination of knowledge graphs and large language models towards more findable and reusable bioinformatics tools.*

Introduction

Ensuring reproducibility in data-driven sciences is critical for the continuous development of open and cumulative sciences. In line with the FAIR principles, this requires for digital scientific resources to be openly accessible and reusable by a wide community of researchers [1,2].

Many registries have been developed to facilitate the discovery and reuse of digital scientific resources. For instance, Zenodo and Dataverse enable the sharing of datasets and increase their discoverability through significant amount of descriptive metadata. These metadata rely on ontologies and generic controlled vocabularies such as Schema.org, DCTerms, or DCAT [3,4,5]. However, the scope of these metadata is generally limited to attribution, citation, or licensing information. They are not sufficient for searching a set of resources annotated with precise concepts, specific to certain scientific disciplines, such as "mobile genetic elements", or "protein-protein interactions".

In the field of life sciences, research communities have developed specialised registries dedicated to training materials (e.g. TeSS [6]), software tools (e.g. bio.tools [7]), or analysis pipelines (e.g. WorkflowHub [8]). These registries rely on EDAM [9], an ontology aimed at improving interoperability in bioinformatics by formally defining the nature and format of data produced and managed, different kinds of data processing, as well as the associated scientific disciplines. This ontology enables,

for example, the retrieval of algorithms dedicated to analyse a specific type of data, as well as relevant training materials. Beyond data findability, semantic indexing allows for the development of computational approaches aimed at assisting scientists in workflow composition [10] or data annotation [11]. These registries, with their growing adoption and extensive collection of resources (e.g., 30k+ bioinformatics entries in bio.tools), are key to address FAIRification challenges.

However, researchers lack insights into the reliability of their annotations. For instance, is a bioinformatics software tool sufficiently annotated? Are the chosen terms precise enough with respect to the terms hierarchy of the domain ontology? To further promote the usage of these domain-specific annotations, we need a detailed quality assessment. In this paper, we address the following question: **What is the quality of semantic annotations associated to bio.tools bioinformatics software ?**

For assessing the quality of annotations, a gold standard is required, but it does not exist yet for bioinformatics software. One approach would then consist in measuring the quality of an annotation according to its rarity: a specific annotation would be less frequent and more informative than a generic annotation, that could be assigned to a large collection of softwares.

Our main contributions are i) a characterisation of the usage of the EDAM ontology when annotating a large collection of bioinformatics software and ii) an evaluation of the specificity of the annotations through the Shannon entropy metric.

Motivating use case

Here we present a small example with two tools to illustrate the EDAM ontology's term hierarchy and its impact on tool search. We selected *Qiime2* [12] and *Vsearch* [13] as two reference bioinformatics tools used in metagenomics data analysis. *Qiime2* is annotated with topics {*Microbial ecology*, *Phylogeny*, *Metatranscriptomics*, *Metagenomics*}, and *Vsearch* with topics {*Metagenomics*, *Sequence analysis*}. The two tools share only one directly assigned annotation {*Metagenomics*}, accounting for 16% of all direct annotations, which is relatively low given their use in the same application domain.

Figure 1 shows the topic annotations for these two tools, as declared in bio.tools, along with inferred annotations from the EDAM class hierarchy. Shared topics between the tools are highlighted by a red border. There are 7 shared topics out of 17 total annotations from the combined sets, representing 41% of the annotations. This highlights the importance of considering inferred annotations when retrieving tool's annotations. This illustrates that some tools have few annotations, and have direct annotations with various levels of precision.

Material and methods

Bio.tools dataset. We leverage the bio.tools registry which now categorizes 30k+ bioinformatics tools using the EDAM ontology. In this work, we rely on the bio.tools RDF metadata available as of January 5, 2025 [6]. Based on the collected metadata, we used SPARQL queries and Python scripts to compute statistics on tool annotations [7]. From the extracted version, there are 30,025 tools described in bio.tools.

6. Available at: https://github.com/research-software-ecosystem/content/blob/master/datasets/bioschemas-dump.ttl
7. Available at: https://github.com/ulysseLeclanche/Abso_bio-tools

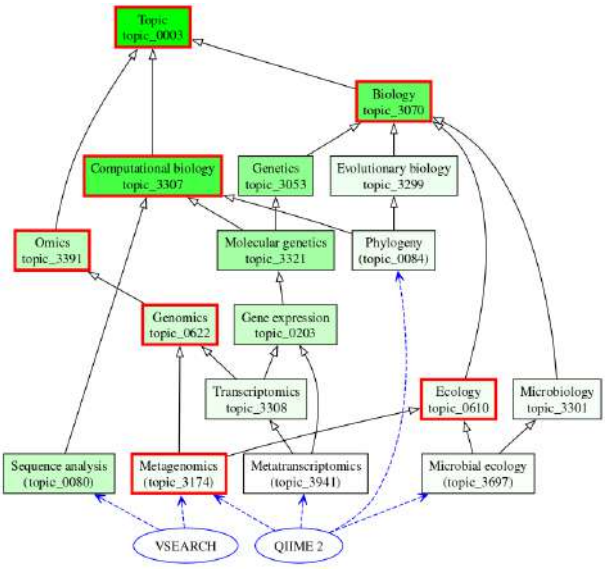


Fig. 1. Direct and inferred EDAM annotations for the scientific topic of Qiime2 and Vsearch, two metagenomics tools. Dotted blue arrows indicate direct topics, while solid black arrows represent inferred topics inherited through the EDAM hierarchy. The topics shared by both tools are highlighted by a red border. Topics colors saturation is proportional to the number of tools annotated by the topics.

EDAM. The EDAM ontology is structured into four main branches covering bioinformatics *Operation*, *Data*, *Format* and *Topic* at different levels of precision. We used EDAM version 1.25, and focused only on *Operation* and *Topic* annotations, the most used annotations in bio.tools (100k+ topics, 70k+ operations, 11k+ data and 10k+ formats). *Data* and *Format* EDAM classes are not well instantiated (used) in the bio.tools registry, even though they are as important as *Topic* and *Operation*. On the extracted version of bio.tools, we observed that branches have a different number of unique classes (331 classes Format, 569 classes Data) than Topics (258 classes Topic) and Operations (527 classes operation). Unfortunately, fewer tools are annotated by elements from these branches (among the 30 025 annotated tools 98.64 % Topics, 94.25 % Operations, 9.64 % Formats, 12.95 % Datas). There are 98,870 topic annotations for 29,616 tools with at least one topic, and 68,886 operation annotations for 28,299 tools with at least one operation.

Entropy. We used Shannon entropy as an information measure to quantify annotation quality [14], as it takes into account annotation rarity and distribution of EDAM annotation. The formula for entropy is : $H = -\sum_{a \in A} p(a) \times \log_2 p(a)$, where p represents the probability of an annotation a occurring. It is defined as the number of tools annotated by a divided by the total number of annotated tools. A low entropy for a tool indicates either that the annotations is general, or that is annotates few tools. A high entropy indicates a balanced distribution in the attribution of annotations to the tool and more specific annotations. For a tool, topic and operation entropy are respectively the sums of their annotation entropy values.

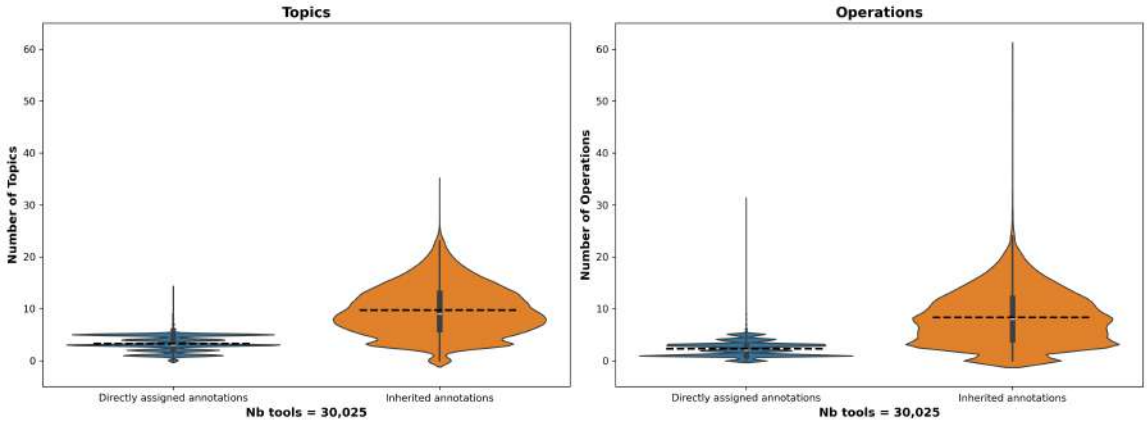


Fig. 2. Distribution of the number of tools in bio.tools according to number of topics and operation. Two conditons are tested: with direct assigned annotations, and with inherited annotations.

Results and Discussion

Topics and operations in bio.tools: basic statistics and curation needs

Basic statistics were computed on the whole bio.tools dataset, comprising 30,025 software descriptions. Figure 2 shows the distribution of the number of tools annotated with a given number of EDAM *topics* or *operations*, considering both directly assigned annotations and inherited annotations. Figure 2 shows a very narrow distribution for directly assigned annotations, with a discrete distribution. This suggests that some annotations are assigned in a very standardized way and are concentrated around a small number of topics and operations. Taking inherited annotations into account yields a wider and more continuous distribution. We observe that the mean number of directly annotated topics is 3.29 $\pm$ 1.45, which is comparable for operations with a means equals to 2.29 $\pm$ 1.43. When inherited annotations (ancestors) are taken into account, the mean number of topics rises to 9.73 $\pm$ 4.71 and to 8.31 $\pm$ 5.12 for operations. The EDAM hierarchy of terms helps to enrich assigned annotations for all tools, taking ancestors into account. These numbers show that the a particular attention should be given to the hierarchy of ontologies classes and not only classes typically used at resource annotation time.

Based on this dataset, we evaluated that 1,965 tools (6.54%) are not annotated with EDAM topic or operation, clearly showing the need for involving user communities to better annotate bioinformatics software. We also computed the number of tools annotated with redundant EDAM classes. For example, the magnet tool has two direct topic annotations: *Protein interactions* and *Molecular interactions, pathways and networks*. These two annotations share the same branch since *Molecular interactions, pathways, and networks* is a subclass of *Protein interactions*. Adding the *Protein interactions* annotation does not provide any additional information, as it is inherited from the ontology. We estimated that 3,405 tools (11.34%) have redundant direct topic annotations, and 2,055 tools (6.84%) have redundant direct operation annotations. This highlights the need for better curation in the tools database. Finally, there are 1,114 deprecated annotations, 54 tools with at least 1 deprecated topic and 347 tools with at least 1 deprecated operation, also highlighting the need for database curation.

Are bioinformatics software annotated with informative enough classes ?

We computed the entropy of tools annotated with EDAM *topics* and *operations*. By only considering direct annotations, the mean *topic* entropy is 0.57 $\pm$ 0.32, but when considering inherited classes, the entropy grows to 2.98 $\pm$ 1.72. We observed the same increase for *operations* with 0.23 $\pm$ 0.18 for direct annotations and 2.01 $\pm$ 1.32 for inherited ones. The entropy of topics for direct and inherited annotations is greater than the entropy of operations. This reflects both a greater diversity in the assignment of topic annotations compared to operation annotations and the size difference of these two branches. We calculated a Pearson correlation coefficient of 0.96 showing a positive correlation between the number of annotations and their entropy. The current entropy measurement can only be increased by adding annotations, even if the annotations are not very informative. However, if we take two tools with a similar number of annotations but different rarity in term annotation, the tool with the most rare terms will have a higher entropy.

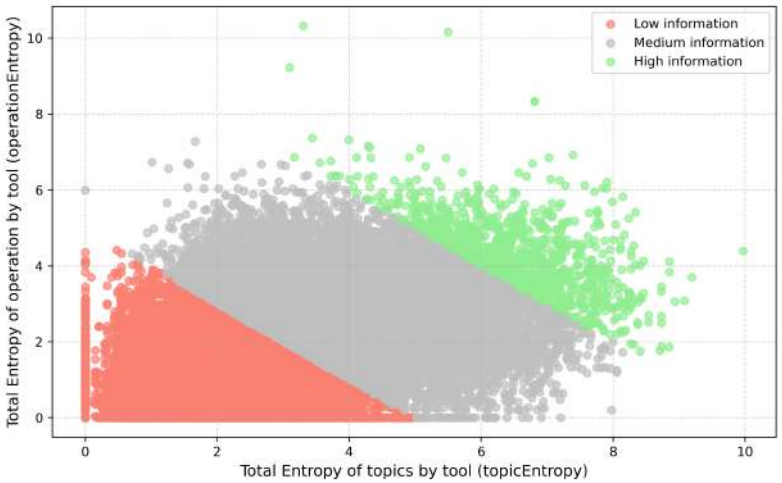


Fig. 3. Distribution of topic and operation entropy for bio.tools software taking into account inherited classes. The total entropy S_e of a tool is the sum of the topic entropy and the operation entropy. Red dots represent tools with low information ($S_e < 5$), gray dots represent tools with medium information $5 <= S_e < 10$, and green dots show tools annotated with highly informative annotations ($S_e >= 10$).

Figure 3 shows how informatively tools are annotated with inherited classes, considering both the *topic* entropy and the *operation* entropy. Tools are grouped into three categories based on an arbitrary threshold on the sum of these two metrics ($max(S_e) = 15.65$). The majority of tools - 29,061 (96.78%) - belongs to the low or medium information categories. A few tools (964) have an entropy sum greater than 10, indicating a high information level for their annotations. 53.05% of tools (in red), i.e. 15,929 tools, are annotated with a low level of information, suggesting that they should be prioritized for database curation.

Increasing the number of annotations has a positive impact on the quality of tool information. The distribution of tools with redundant annotations is similar in each group, with 2,369 (14.87%) redundant tools in the low information level, 2,193 (16.70%) in the medium level and 86 (8.92%) in the high level. However, among the top 10 tools with the highest entropy sum, 60% of tools have redundant

annotations. Redundant annotations should be removed, as they artificially increase entropy. A metric that penalizes, more than entropy, annotation generality and redundancy would be interesting.

Conclusion

In this work, we have shown that considering inherited annotations from the ontology increases the number of annotations for topics and operations in the whole set of tools, making the tools more searchable and reusable. To quantify annotation quality, we used Shannon entropy, which takes into account annotation rarity and the distribution of EDAM annotations. This measure enabled us to compare the annotation quality of the tools with each other, and to identify a set of 15,929 tools (53.05%) with a low level of information annotations. This set of tools should be prioritized for future database curation activities. The main limit of our approach is the lack of ground-truth to assess the accuracy of annotations, Shanon entropy assess the rarity of annotations, which does not mean that they are correct. To address this issue, we are currently working with bioinformatics experts to define a reference dataset of highly curated annotations. Through this study, we have also seen the impact of EDAM ontology evolution on annotation quality, with the identification of redundant or obsolete annotations. The more limited usage of the *Format* and *Data* branches suggest that even if they have a smaller impact on tools annotation, the margin for improvement is also higher than for *Topic* or *Operation*.

This opens for new research directions we will pursue as future works. We are currently working on implementing more suited metrics to better assess the quality of EDAM annotations. To support curation tasks, we aim at combining large language models and knowledge graphs [15] as a means to suggest more informative annotations, or to identify possibly missing classes in the ontology. Evaluating the benefits of enriched annotations and new EDAM classes requires an expert-approved gold standard. Identifying and quantifying missing classes in ontologies remains a challenge. Overrepresentation of certain domains in ontologies, as seen in Gene Ontology, can stem from factors like research focus, annotation specificity, or ontology structure. This imbalance complicates entropy normalization and requires expert analysis, as annotation rarity is not equivalent to accuracy. Although this work is grounded to bio.tools, it aims at being generalized to other application domains also using ontologies and registries for annotating and sharing FAIR digital resources.

Availability and Implementation

All the code for extracting metadata from the RDF schema, creating article figures and calculating tool annotation statistics is available on the following github repository: https://github.com/ulysseLeclanche/Abso_bio-tools.

Funding information

This work is supported by the Agence Nationale de la Recherche under the France 2030 program, ANR-22-PESN-0007 ShareFAIR.

References

[1] Kinkade D, Shepherd A. Geoscience data publication: Practices and perspectives on enabling the FAIR guiding principles. Geoscience Data Journal. 2022;9(1):177-86.

[2] Top J, Janssen S, Boogaard H, Knapen R, Şimşek-Şenel G. Cultivating FAIR principles for agri-food data. Computers and Electronics in Agriculture. 2022;196:106909.

[3] Guha RV, Brickley D, Macbeth S. Schema.org: Evolution of Structured Data on the Web. Queue. 2015;13:10 37. Available from: https://api.semanticscholar.org/CorpusID:27038003.

[4] Weibel SL, Kunze JA, Lagoze C, Wolf M. Dublin Core Metadata for Resource Discovery. RFC. 1998;2413:1-8. Available from: https://api.semanticscholar.org/CorpusID:43249830.

[5] Archer P. Maali F, Erickson J, editors. Data Catalog Vocabulary (DCAT) (W3C Recommendation); 2014. Online. Available from: https://www.w3.org/TR/vocab-dcat/.

[6] Beard N, Bacall F, Nenadic A, Thurston M, Goble CA, Sansone SA, et al. TeSS: a platform for discovering life-science training opportunities. Bioinformatics. 2020 02;36(10):3290-1. Available from: https://doi.org/10.1093/bioinformatics/btaa047.

[7] Ison JC, Ienasescu H, Chmura P, Rydza E, Ménager H, Kala M, et al. The bio.tools registry of software tools and data resources for the life sciences. Genome Biology. 2019;20. Available from: https://api.semanticscholar.org/CorpusID:199538589.

[8] Gustafsson J, Wilkinson SR, Bacall F, Pireddu L, Soiland-Reyes S, Leo S, et al. WorkflowHub: a registry for computational workflows. ArXiv. 2024;abs/2410.06941. Available from: https://api.semanticscholar.org/CorpusID:273228446.

[9] Ison JC, Kala M, Jonassen I, Bolser DM, Uludag M, McWilliam H, et al. EDAM: an ontology of bioinformatics operations, types of data and identifiers, topics and formats. Bioinformatics. 2013;29:1325 1332. Available from: https://api.semanticscholar.org/CorpusID:1626822.

[10] Kasalica V, Schwämmle V, Palmblad M, Ison J, Lamprecht A. APE in the Wild: Automated Exploration of Proteomics Workflows in the bio.tools Registry. Journal of proteome research. 2021;20(4):2157–2165. Publisher Copyright: © 2020 American Chemical Society. All rights reserved.

[11] Gaignard A, Skaf-Molli H, Belhajjame K. Findable and reusable workflow data products: A genomic workflow case study. Semantic Web – Interoperability, Usability, Applicability. 2020 May:1-13. Available from: https://hal.science/hal-02903805.

[12] Bolyen E, Rideout JR, Dillon MR, Bokulich NA, Abnet CC, Al-Ghalith GA, et al. QIIME 2: Reproducible, interactive, scalable, and extensible microbiome data science; 2018. .

[13] Rognes T, Flouri T, Nichols B, Quince C, Mahé F. VSEARCH: a versatile open source tool for metagenomics. PeerJ. 2016;4.

[14] Shannon CE. A mathematical theory of communication. The Bell System Technical Journal. 1948;27(3):379-423.

[15] Gilbert S, Kather JN, Hogan A. Augmented non-hallucinating large language models as medical information curators. NPJ digital medicine. 2024;7(1):100.

A decade of strengthening bioinformatics in West Africa: HPC infrastructure, training, and scientific collaboration

Ezechiel B. Tibiri[1,2], Christine Dubreuil- Tranchant[3*], Romaric K. Nanema[4], Fidèle Tiendrebeogo[1,2], Justin S. Pita[2]

[1]Laboratoire de Virologie et de Biotechnologies Végétales, Institut de l'Environnement et de Recherches Agricoles (CNRST/INERA), Ouagadougou, Burkina Faso

[2]Central and West African Virus Epidemiology (WAVE), Pôle scientifique et d'innovation de Bingerville, Université Félix Houphouët-Boigny (UFHB), Bingerville, Côte d'Ivoire

[3]DIADE, University of Montpellier, CIRAD, IRD, 911 Avenue Agropolis, 34934 Montpellier Cedex 5, France

[*]South Green Bioinformatics Platform, Bioversity, CIRAD, INRA, IRD, Montpellier, France Montpellier, France

[4]Genetic and Plant Breeding Team (EGAP), Biosciences Laboratory, Doctoral School of Science and Technology, Joseph KI-ZERBO University, Burkina Faso

Corresponding author:

Abstract

Since 2014, a collaborative and interdisciplinary dynamic has emerged in West Africa to build lasting capacities in bioinformatics. Driven by the growing need to analyze locally produced sequencing data, this initiative has led to the development of regional infrastructures and training programs through strong partnerships between academic and research institutions, including Joseph KI-ZERBO University (UJKZ), INERA, IRD, and the LMI PathoBios. Key milestones include the establishment of bioinformatics platforms in Ouagadougou (Burkina Faso) and, more recently, in Bingerville (Côte d'Ivoire) within the WAVE-CI framework.

These platforms have served as training hubs, enabling a wide range of hands-on and theoretical training—from basic GNU/Linux usage to advanced metagenomics data analysis. A major achievement of this initiative is the launch of the **International Certificate in Bioinformatics and Genomics (CIBiG)** in 2023–2024. This intensive program combines 154 hours of in-person courses and practical sessions with laboratory work, project-based tutoring, and personalized coaching. It covers the entire data lifecycle, from sequencing using Oxford Nanopore Technologies (ONT) to data analysis workflows including assembly, annotation, SNP detection, phylogenetics, and transcriptomic analyses.

Anchored in a participatory and inclusive model, CIBiG addresses two main objectives: (1) strengthening local expertise in bioinformatics applied to agriculture and health, and (2) structuring a regional community of practice. The program is supported by committed institutional stakeholders (UJKZ, IRD, WAVE), a broad network of trainers, and a strong ambition to sustain the initiative through curriculum reforms, long-term funding strategies, and regional thematic working groups.

This paper presents a ten-year retrospective on capacity-building activities, the impact of the co-constructed training programs, the pedagogical innovations used (e.g., JupyterBook, Slack, supervised internships), and the perspectives for scaling up this pioneering experience in West Africa.

Madbot, a metadata and data brokering online tool to ensure the adoption of standards and FAIR principals in an open science context

Laurent BOURI*[1], Imane MESSAK*[1], Baptiste ROUSSEAU*[1], Anakim GUALDONI[1], Elora VIGO[1], Matéo HIRIART[2], Nadia GOUÉ[1,2], Julien SEILER[#1], Thomas DENECKER[#1]

1 IFB-core, Institut Français de Bioinformatique (IFB), CNRS, INSERM, INRAE, CEA, 91057 Evry, France

2 Université Clermont Auvergne, Plateforme AuBi, Mésocentre Clermont-Auvergne, F-63000 Clermont-Ferrand, France

Corresponding Author: thomas.denecker@france-bioinformatique.fr

* Co first authors # Co last authors

Keywords

Data brokering, Open Science, Online tool, Metadata, Data submission

Abstract

Madbot is a tool designed to help researchers manage and share their scientific data more easily. As research data continues to grow in volume, it becomes harder to ensure that data is accessible, reusable, and easy to understand. While other tools exist to help with parts of this process, they often lack automation, standardization, or flexibility. Madbot solves these issues by providing a simple and comprehensive solution that follows international data standards, making it easier for researchers to publish their data. It automates much of the work involved in organizing and describing data, which saves time and effort for researchers. Madbot also helps ensure that data is described correctly and consistently, following well-established standards. This makes it easier for others to find and use the data in the future. The tool connects to various global platforms like Zenodo and ENA (European Nucleotide archive), allowing researchers to submit their data directly to these repositories without hassle. Madbot's easy-to-use interface allows users to interact with the system even if they don't have technical expertise. Behind the scenes, the tool keeps everything organized, automatically checks for mistakes, and helps researchers create accurate and high-quality metadata. Madbot's architecture is designed to be easily extensible, enabling integration with various data storage solutions, data repositories, and metadata standards. This flexibility allows researchers to adapt the tool to their specific needs, ensuring seamless interoperability with different research infrastructure. By simplifying the process of submitting research data, Madbot encourages researchers to adopt open science principles, making their work more accessible to others. In the end, Madbot helps reduce the

barriers to sharing research data and makes it easier for scientists to contribute to the global scientific community.

Introduction

The exponential growth of data production across scientific disciplines presents a major challenge for metadata and data management. To ensure long-term usability, data must be well-documented, discoverable, and interoperable. However, researchers often perceive metadata creation as tedious, leading to incomplete descriptions that hinder data reuse and integration. Additionally, the heterogeneity of standards complicates interoperability across disciplines and repositories. International platforms like the European Nucleotide Archive (1), Zenodo (2) and Dataverse (3) ensure data preservation and accessibility, but their submission processes can be complex and time-consuming. This additional workload discourages researchers, limiting the impact of open data initiatives. There is a pressing need for a tool that supports researchers in managing data and metadata while streamlining submission to repositories. Such a solution should facilitate metadata enrichment, automate submissions, and improve data visibility while reinforcing the adoption of FAIR principles (Findable, Accessible, Interoperable, Reusable).

Existing tools address parts of this challenge but lack comprehensive solutions (Tab. 1). Data managers like Onedata (4) or iRODS (5) do not handle metadata submission, while metadata managers such as FAIRdom Seek (6) do not integrate data storage. Automated submission tools like MARS (7) require specific formats that researchers may not use. All-in-one tools like athENA (8), METAGENOTE (9) or Maggot (10) are often restricted to specific disciplines or repositories with less stringent metadata requirements.

To bridge this gap, we developed Madbot, a Metadata And Data Brokering Online Tool (Fig. 1). Our tool offers a hierarchical structure for projects, which facilitates the organization and description of research data. However, Madbot is not a data storage tool. Research data remains in its usual storage location, and Madbot accesses it through a collection of connectors. Once data is associated with a project, users have access to an interactive dashboard that provides an overview of the data's status and accessibility, helping to optimize their management. These data can be described using metadata fields from Madbot’s own referential or from internationally recognized standards widely adopted by the scientific community. Implementing a metadata referential ensures standardized and consistent descriptions, making interoperability between data sources and submissions to international repositories easier. Automatic quality control further ensures metadata compliance with standards, thereby facilitating data sharing. Finally, Madbot enables the publication of both metadata and data to various repositories, whether they are general or specialized. By integrating these features, we aim to reduce researcher's workload, improve metadata quality, and promote the adoption of FAIR principles. Each of these points will be discussed in more detail in the results section.

1. Methods

1.1. Architecture of the tool

The tool follows a modular, scalable architecture, integrating frontend and backend components (Fig. 2). The backend is built with Django, serving as the core API and interacting with a SQL database for data storage. Celery (11) manages asynchronous tasks, with Redis (12) as the message broker. A key feature is the integration of HashiCorp Vault (13), which securely stores user credentials required for authentication with external repositories (Galaxy (14), NAS, cluster...), enabling dynamic data link maintenance. The Django application uses the ASGI (15) protocol with WebSocket support, and the Green Unicorn (16) server efficiently handles multiple workers. On the frontend, Nuxt.js (17) and Vuetify (18) create a dynamic, user-friendly interface. The development lifecycle is automated with GitLab CI/CD pipelines, for both frontend and backend. These pipelines ensure code quality via Ruff (19) or Eslint (20), check for security vulnerabilities with Safety (21) and bandit (22), build the backend (Python) and frontend (JavaScript), and store the package in a registry. They also generate Docker (23) images, create a new release with semantic release, and automatically update the changelog. Deployment is containerized and orchestrated for high availability, scalability, and security.

1.2. Design considerations

For Madbot's development, special attention was given to its architecture to ensure a modular, accessible, and user-friendly solution. On the backend, the goal was to create a simple, well-documented, and reusable API (Application Programming Interface), enabling seamless integration with potential clients like automated scripts or Laboratory Information Management Systems (LIMS). The API follows the Open API Specification v3 (24) and is fully documented with Swagger (25), ensuring easy understanding for developers. Madbot features its own identity provider built using the Django OAuth Toolkit (26), enabling connection management via OAuth2. Additionally, authentication can be delegated to external IDPs through Django Allauth (27), supporting any OIDC and SAML providers such as ORCID (28), LS login (29) or other identity federation. On the client side, the focus was on creating a clear, intuitive interface to simplify metadata management, particularly during submissions to international repositories. In this regard, Madbot's interface is intended for all users (researchers, IT, …), offering an accessible environment that doesn't require advanced technical skills. Features include contextual input assistance, advanced search, real-time validation, and a guided tour system to help users navigate the tool.

1.3. Metadata in JSON Schema Format

Madbot uses the JSON Schema (30) format to structure and validate metadata, ensuring rigorous formalization and interoperability with other tools and infrastructures. This standardized format defines data types, constraints, and relationships between object elements, enabling automatic

metadata validation. With this approach, Madbot offers an interactive interface that guides users. Additionally, the structured format simplifies converting and transmitting metadata to external repositories while ensuring compliance with their specific requirements. By adopting JSON Schema, Madbot facilitates metadata models evolution, allowing administrators and developers to introduce new fields or modify validation constraints without requiring major changes to the core application.

2. Results

2.1. Workspaces and nodes organization

Madbot provides a customizable workspace system for users to create dedicated workspaces, invite collaborators, and manage permissions with granular access control. This ensures secure collaboration and clear separation of projects and contributors, maintaining structure and independence. Each workspace includes a dashboard offering insights into data, activities, updates, submission status, and metadata evolution, enhancing data management efficiency and research integrity. To describe projects, Madbot uses tree structures based on nodes, each representing a specific part of the research project, including titles, descriptions, and data associations. Additionally, each node can contain metadata, biological samples and subnodes. This structure follows the ISA framework (Investigation, Study, Assay) (31), organizing research hierarchically and improves data traceability and interoperability, especially in life sciences. While the model's flexibility can cause inconsistencies across projects, it offers researchers freedom while maintaining structure. Future plans include adding other hierarchical structures to improve project management.

2.2. Connectors

Madbot's connectors are plugins designed to enhance its functionalities and facilitate interactions with data sources and submission repositories (Tab. 2). Currently, there are two types of connectors: data connectors and submission connectors. Data connectors generate a link which locates a file owned by the user and enables its download. Currently, Madbot includes two operational data connectors: Galaxy and SSHFS, with future plans for Omero (32), DeepOmics (33), GitHub (34), GitLab (35), Nextcloud (36), ENA, Zenodo, and Dataverse. Submission connectors streamline the collection and organization of metadata for submission to remote databases. As of now, Madbot supports Zenodo and ENA, with plans to add connectors for PRIDE (37) and Dataverse. Special attention has been given to designing a connector API for optimal flexibility, allowing quick integration of new connectors and external contributions. Several extensions are planned for future developments.

2.3. Dashboard

Research data production often involves multiple contributors, making it difficult to track data location and metadata provenance, especially in integrative bioinformatics. Madbot addresses this by

providing a comprehensive dashboard that offers real-time insights into data storage, accessibility, and metadata quality. Users can monitor submission progress and evaluate metadata completeness with visual indicators and automated validation checks. The platform streamlines the submission workflow, automating metadata curation, data retrieval, and validation before submission to ensure compliance with repository standards and FAIR principles. It securely retrieves datasets, checks data integrity, and facilitates submissions.

2.4. The metadata reference

As of mid-2025, the FAIRsharing (38) catalog lists over 1,830 metadata standards. Each international repository offers its own metadata fields reference framework and makes distinct choices regarding data format control or the adoption of ontologies. For instance, fields such as a list of collaborators or associated publications are structured differently across major repositories like ENA, Zenodo, or Dataverse. When it comes to more technical details, such as a sequencing method or sample collection procedure, the disparities between repositories become even more pronounced. To address this, Madbot has designed a central metadata reference framework that integrates widely adopted scientific frameworks. This framework provides a standardized structure, ensuring that each submission connector can interpret metadata correctly. Madbot maps these metadata fields to the corresponding fields in submission repositories, helping users understand how their data aligns with each platform's requirement. Through the Madbot API, external fields and their mapping to the central reference framework are made accessible, ensuring consistency in metadata handling.

2.5. Streamlining data description with inheritable and connected metadata

Madbot reduces metadata entry effort by implementing inheritance in its tree structure, allowing higher-level metadata to automatically propagate to related datasets. This eliminates redundancy while ensuring consistency, with the option for researchers to override inherited values. This feature is especially useful for collaborative projects and submissions to multiple repositories, enabling platform-specific customization. Madbot also enhances metadata by integrating with authoritative external sources like ROR (39), ORCID, DOI (40), PubMed (41), and NCBI Taxon ID (42), improving the quality and discoverability of datasets. It supports multiple repositories and metadata standards while maintaining versioning and traceability, facilitating compliance and interoperability.

2.6. Data brokering

Once the user has linked data and provided the necessary metadata, a submission can be initiated to either a thematic repository (e.g., ENA for sequencing data) or a general-purpose repository (e.g., Zenodo). Madbot ensures metadata interoperability by converting it into the required format for the target repository and verifying data integrity and accessibility. The submission connectors organize

metadata into batches, which are automatically updated when data is added or removed. Users can adjust metadata values as needed. This flexible architecture accommodates both simple (Zenodo) and complex (ENA) metadata structures. The system retrieves data, generates necessary files (e.g., manifests, validation reports), and submits asynchronously for efficiency. Upon successful submission, Madbot stores persistent identifiers (e.g., DOIs, accession numbers) for easy referencing, streamlining the process and ensuring repository compliance with minimal effort for researchers.

3. Discussion & conclusion

Madbot is a powerful tool for metadata and data management, offering several key advantages in research and open science environments. Its flexible architecture simplifies data flow, enabling the seamless collection, integration, and management of metadata from various sources. The intuitive dashboard enhances decision-making by providing real-time visualization of data and metadata. Additionally, Madbot's connected metadata system ensures consistency and traceability across different platforms, crucial for transparency and reproducibility in research. Its hyper-flexible connector system enhances compatibility and scalability by easily adapting to new integration needs.

Despite these strengths, Madbot has some limitations that need to be addressed. The current number of available connectors is limited, which may restrict its use in specific contexts (Table 1). Additionally, the tool has not been tested under heavy load, which could pose challenges in large-scale environments. To address these issues, future improvements will focus on expanding the connector ecosystem, including new connectors for metadata import/export, data management plans, and specialized tools like iRODS. Additionally, repository feedback will be processed and integrated into the dashboard, enhancing tracking and troubleshooting capabilities. These updates will help meet the needs of the scientific community, particularly in fields like biology. Madbot's current architecture, while effective, presents some operational and security challenges. Its reliance on directly integrated connectors limits access to internal storage systems and creates potential security risks due to broad access permissions. To resolve these issues, Madbot's connector layer will be separated into a standalone Connector API, improving security and flexibility. This change will also allow users to deploy the Connector API locally, enabling access to private storage systems while maintaining strict compartmentalization. As part of ongoing developments, Madbot will soon integrate a feature that automatically generates metadata sets based on node descriptions, using advanced text generation tools from Large Language Models. These updates will further enhance Madbot's usability, security, and overall functionality, paving the way for broader adoption in research data management.

Images

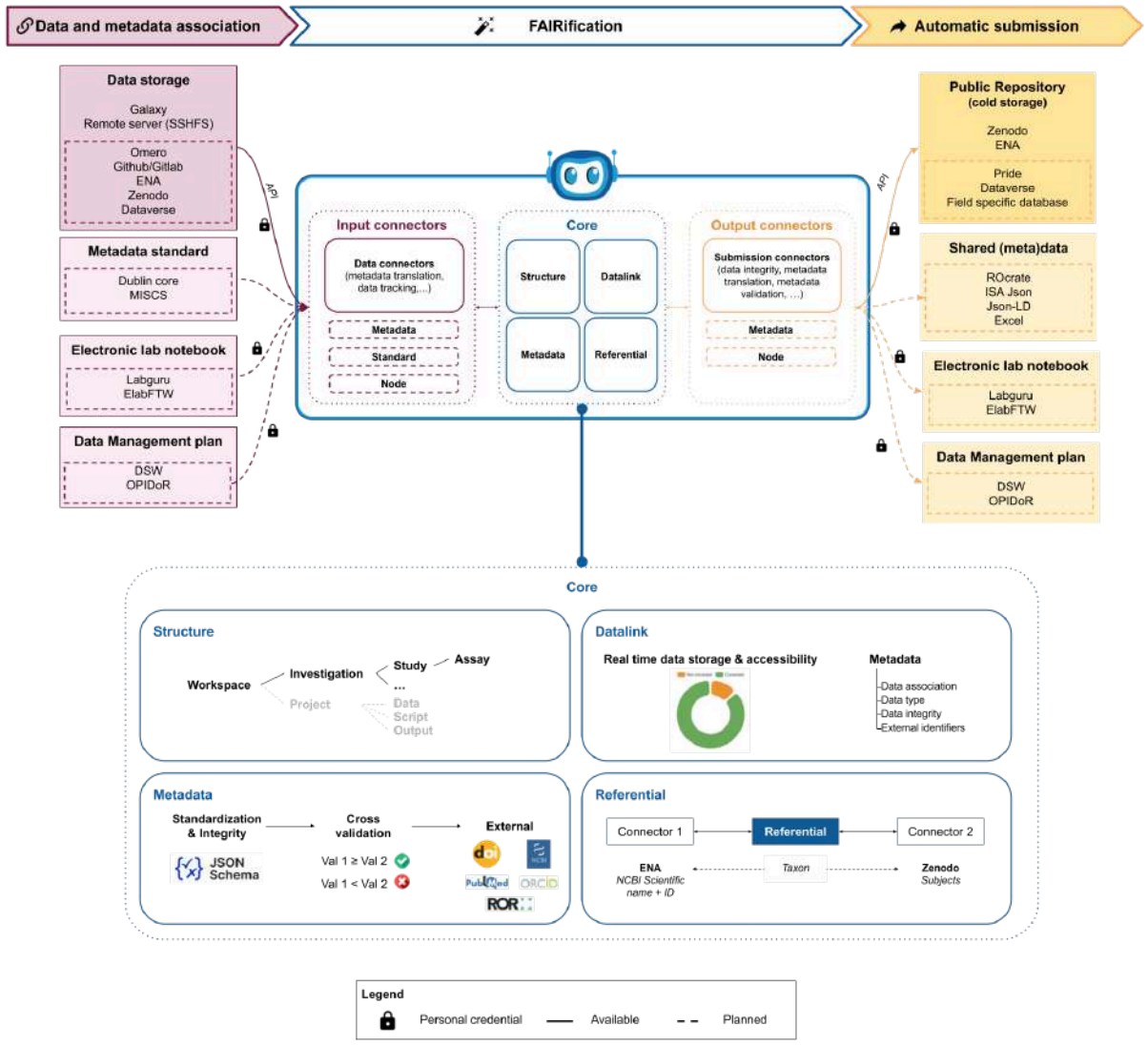


Fig. 1: *Madbot architecture and workflow for data and metadata management in an integrative bioinformatics context. The diagram illustrates the integration of various data sources (data storage, electronic lab notebooks, data management plans) through input connectors. Madbot standardizes, validates, and enriches metadata using external repositories and recognized standards to ensure integrity and FAIR compliance. Output connectors automate submission to public and shared repositories, promoting data accessibility and traceability. The hierarchical structure of information and validation indicators provide real-time monitoring of stored data and associated metadata.*

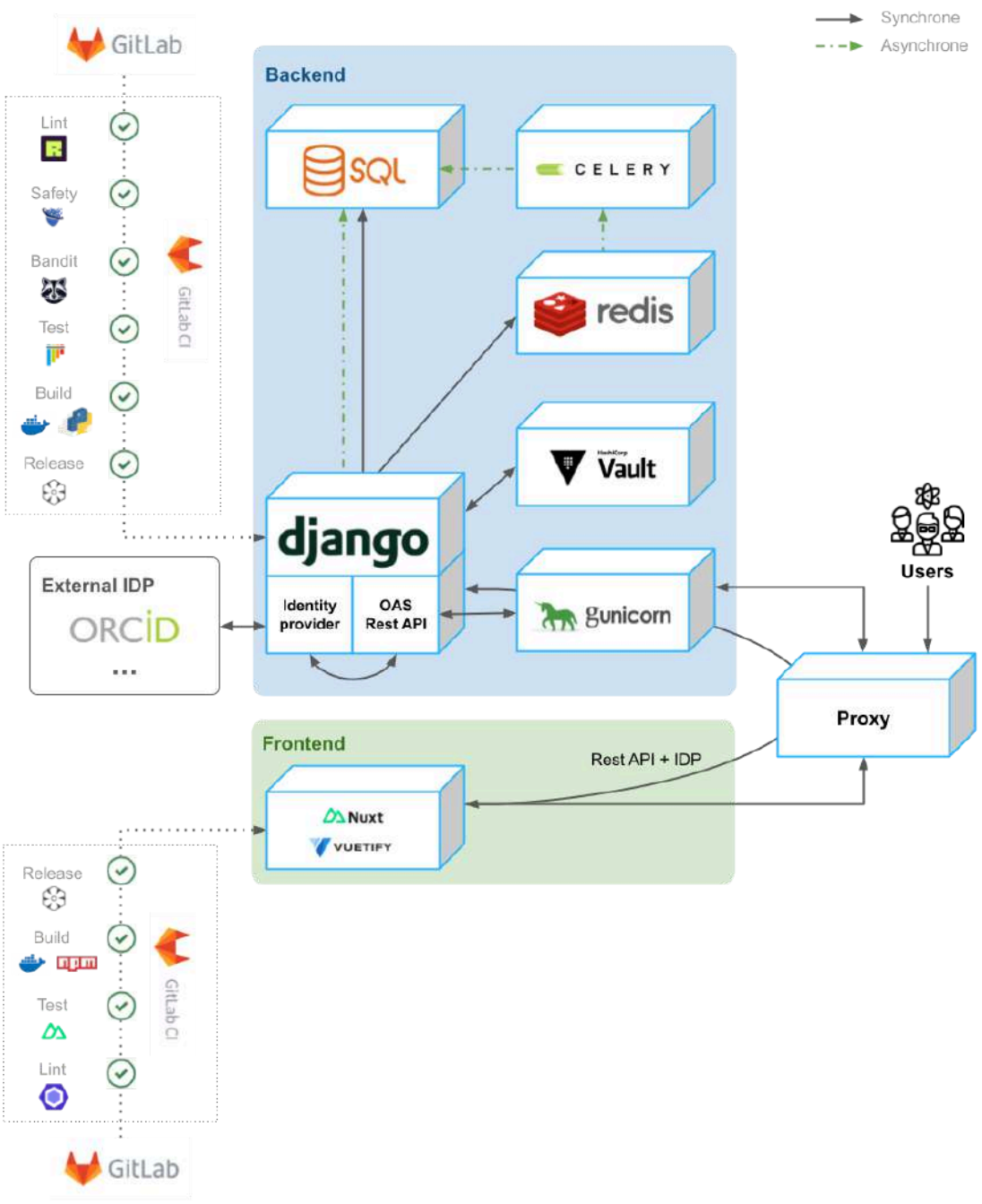


Fig. 2: *Overview of the Madbot technical architecture, integrating backend and frontend components for secure and efficient data management. The backend, based on Django, handles authentication, data orchestration, and metadata management. Asynchronous tasks are managed via Celery and Redis, while Vault ensures secure handling of sensitive information. The SQL database stores structured data, and Gunicorn manages HTTP requests. The frontend, built with Nuxt and Vuetify, interacts with the backend through REST API and IDP (Identity Provider) authentication using ORCID. The entire system is deployed and maintained using GitLab CI/CD for continuous integration and deployment.*

Features	athENA	FAIRdom Seek	iRODS	METAGE NOTE	Maggot	MARS	Onedata	Madbot
Data Management Life Cycle								
Capture, Create & Collect								
Import Data	-	Yes	Yes	Yes	Yes	-	Yes	No
Organize & Store								
Hierarchical Data Organization	No	Yes (ISA)	Yes (collections)	Yes (ENA)	Yes	No	Yes (Datasets)	Yes (ISA, …)
Act as a storage backend	-	Yes	Yes	No	Yes	-	Yes	No
Access to remote storage	-	No	No	No	Yes	-	Yes	Yes
Use & Analyse								
Data Analysis	No	Yes	No	No	No	No	No	No
Data Visualization	No	No	Yes	No	Yes	No	No	Planned
Export Data	No	Yes	Yes	No	Yes	No	Yes	No
Share								
Submission to Repositories	Yes (ENA)	Yes (Zenodo)	No	Yes (NCBI's SRA)	Yes (Zenodo, Dataverse)	Yes (ENA, BioSamples)	No	Yes (Zenodo, ENA)
Seamless connection with new repositories	No	No	No	No	No	Yes	No	Yes
Interoperability with Multiple Submission Repositories	No	Yes	No	Yes	Yes (Zenodo, Dataverse)	No	No	Yes
Sample Management								
Sample Creation	Yes	Yes	No	Yes	No	No	No	Yes
Sample Management	Yes	Yes	No	Yes	No	No	No	Yes
Sample associated with data	Yes	Yes	No	Yes	No	No	No	Yes

**A hyphen represents "Not applicable" or "Unspecified"*

***The data in this table is accurate as of the time of writing and to the best of our understanding within the constraints of the access to relevant tools.*

****All referenced tools are cited within the article.*

Features	athENA	FAIRdom Seek	iRODS	METAGE NOTE	Maggot	MARS	Onedata	Madbot
Metadata Management								
Metadata Creation	Yes	Yes	Yes	Yes	Yes	No	Yes	Yes
Metadata Management	Yes	Yes	Yes	Yes	Yes	No	Yes	Yes
Association of metadata to hierarchical elements	-	-	-	-	-	-	-	Yes
Association of metadata to samples	Yes	Yes	No	Yes	No	No	No	Yes
Association of metadata to data	Yes	Yes	Yes	Yes	Yes	No	Yes	Yes
Internal Metadata fields referential	No	Yes	No	No	Yes	No	Yes	Yes
Support External Metadata fields referential	Yes (ENA)	No	No	Yes (SRA)	Yes (Zenodo, Dataverse)	Yes (ENA)	No	Yes (submission & standard connectors)
Seamless integration of external metadata referential	No	No	No	No	No	Yes	No	Yes
Mapping between referential	No	No	No	No	Yes	No	No	Yes
Integration with External Authorities	Yes (NCBI taxon)	Yes	No	Yes (NCBI taxon)	No	Yes	No	Yes (ROR, ORCID, DOI, PubMed, Ncbi Taxon)
Import Metadata	No	Yes	Yes	Yes	No	No	No	Planned
Export Metadata	Yes (Excel sheet)	Yes (ISA-JSON format)	Yes (JSON)	Yes (Excel sheet)	Yes (Zenodo, Dataverse, JSON-LD)	No	No	Planned

**A hyphen represents "Not applicable" or "Unspecified"*

***The data in this table is accurate as of the time of writing and to the best of our understanding within the constraints of the access to relevant tools.*

****All referenced tools are cited within the article.*

Features	athENA	FAIRdom Seek	iRODS	METAGE NOTE	Maggot	MARS	Onedata	Madbot
Metadata Visualization	No	Yes	No	No	No	No	No	Planned
Metadata Inheritance Mechanism	No	No	No	No	No	No	No	Yes
Validation of metadata upon entry	Yes	-	No	Yes	Yes	No	No	Yes
Automated Metadata translation between referential	No	No	No	No	Yes	Yes	No	Yes
Validation of metadata before submission	Yes	-	No	Yes	Yes	Yes	No	Yes
Standard traceability and versioning	No	No	No	No	No	No	No	Yes
Usability & Collaboration								
Graphical User Interface	No	Yes	No	Yes	Yes	No	Yes	Yes
Documentation of GUI	-	Yes	-	Yes	Yes	-	Yes	Planned
Command Line Interface	Yes	No	Yes	No	No	Yes	No	No
Documentation of CLI	Yes	Yes	No	-	-	Yes	-	-
Programmatic Interface	No	Yes	Yes	No	No	No	Yes	Yes
Documentation of programmatic interface	-	Yes	Yes	-	-	-	Yes	Yes
Users and roles management	No	Yes	Yes	No	Yes	No	Yes (Groups)	Yes
Collaboration Features	No	Yes (Project hubs)	Yes	No	Yes	No	No (Spaces)	Yes (All levels)

**A hyphen represents "Not applicable" or "Unspecified"*

***The data in this table is accurate as of the time of writing and to the best of our understanding within the constraints of the access to relevant tools.*

****All referenced tools are cited within the article.*

Features	athENA	FAIRdom Seek	iRODS	METAGE NOTE	Maggot	MARS	Onedata	Madbot
Project FAIRness								
Data FAIR Principles Support	Yes	Yes	No	Yes	Yes	Yes	No	Yes
Tool FAIR Principles Support	Yes	Yes	No	Yes	Yes	Yes	No	Yes
Open-Source Project	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Security & Compliance								
Secure Credential Management	-	-	-	-	-	-	-	Yes (HashiCorp Vault)
Various Authentication Methods	No	Yes (Github, LS login)	Yes	Yes (OIDC)	-	No	Yes (EGI, Google)	Yes (OIDC, SAML, CAS)
Architecture and Technical Features								
Modular Architecture	No	No	Yes	No	No	Yes	No	Yes
Scalability and High Availability	Yes	Yes	Yes	Yes	Yes	No	Yes	No

**A hyphen represents "Not applicable" or "Unspecified"*

***The data in this table is accurate as of the time of writing and to the best of our understanding within the constraints of the access to relevant tools.*

****All referenced tools are cited within the article.*

Table 1: *Feature comparison of Madbot and existing tools.*

	Data connector	Submission connector	Node importer	Node exporter	Metadata Importer	Metadata exporter
Galaxy	Ready	-	-	-	Planned	-
SSHFS	Ready	-	-	-	Planned	-
ENA	Planned	Ready	Planned	-	Planned	-
Zenodo	Planned	Ready	-	-	-	-
Dataverse	Planned	Planned	-	-	-	-
iRODS	In progress	-	-	-	-	-
OMERO	In progress	-	Planned	-	Planned	Planned
Labguru (43)	-	-	In progress	In progress	-	-
ElabFTW (44)	-	-	Planned	Planned	-	-
DeepOmics	In progress	-	In progress	-	In progress	-
Seafile (45)	In progress	-	-	-	Planned	-
DMP OPIDoR (46)	-	-	Planned	Planned	Planned	Planned
DSW (47)	-	-	Planned	Planned	Planned	Planned
ROCrate (48)	-	-	Planned	Planned	Planned	Planned
JSON-LD (49)	-	-	-	-	Planned	Planned
GitLab	Planned	-	-	-	-	-
GitHub	Planned	-	-	-	-	-
Amazon S3 (50)	Planned	-	-	-	-	-

**A hyphen represents "Not applicable"*

***All non-referenced tools are cited within the article*

Table 2: *Current and prospective connectors of Madbot.*

Availability and Implementation

All links were verified on the 2025/03/12.

- Project name: Madbot
- Project homepage: https://ifb-elixirfr.gitlab.io/madbot/madbot-doc/
- Project code repository:
    - Back: https://gitlab.com/ifb-elixirfr/madbot
    - Client: https://gitlab.com/ifb-elixirfr/madbot/madbot-client
    - Documentation: https://ifb-elixirfr.gitlab.io/madbot/madbot-doc/
- Operating system(s): Platform independent
- License: BSD 3-Clause License
- Programming languages (all versions are available in different repositories):
    - Back: Django Python
    - Front: Vue3, Typescript, Nuxt
    - Doc: Mkdoc & Material
- Other repositories:
    - Software Heritage:
        - API: https://archive.softwareheritage.org/swh:1:dir:41825e65674d7d5de78f43365a36aec4b76c8ab3;origin=https://gitlab.com/ifb-elixirfr/madbot/madbot-api;visit=swh:1:snp:fa90259e758590d6093a2346024d1123d636514d;anchor=swh:1:rev:05cda75f6e7aa373cf04fb6882f600aa4bf35a8e
        - Client: https://archive.softwareheritage.org/swh:1:dir:ab8450f3b5a4bfd4ec6ccc9c5a94e3e92172e7fe;origin=https://gitlab.com/ifb-elixirfr/madbot/madbot-client;visit=swh:1:snp:c66e0b05732295ecd8b8be82fe1ea97ae460080d;anchor=swh:1:rev:edbc02d9454a1a649611772e6ebefd760e8751f8
        - Doc: https://archive.softwareheritage.org/swh:1:dir:b78f8317d203d4d938a4443fedc4e69de0740ae8;origin=https://gitlab.com/ifb-elixirfr/madbot/madbot-doc;visit=swh:1:snp:05f27157f0281e911596c7f200e0cc8d1037f357;anchor=swh:1:rev:400cc56fcaa7a4bdcdd7ed8256e13699683d4d17

Author Contributions

Laurent BOURI: Conceptualization (equal); Software (lead); Writing – Review & Editing (equal). **Imane MESSAK**: Conceptualization (equal), Software (lead) and Writing – Review & Editing (equal). **Baptiste ROUSSEAU**: Conceptualization (equal), Software (lead) and Writing – Review & Editing (equal). **Anakim GUALDONI**: Conceptualization (equal), Software (lead) and Writing – Review & Editing (equal). **Elora VIGO**: Conceptualization (equal), Software (lead) and Writing – Review & Editing (equal). **Matéo HIRIART**: Software (supporting). **Nadia GOUÉ**: Supervision (supporting); Writing – Review & Editing (equal). **Julien SEILER**: Conceptualization (lead); Project administration (lead); Software (lead); Supervision (lead); Writing – Review & Editing (equal). **Thomas DENECKER**: Conceptualization (lead); Project administration (lead); Software (lead); Supervision (lead); Writing – Review & Editing (equal).

The authors used generative AI to improve the readability and language of their own writing.

Acknowledgements

We would like to express our gratitude to our working group for their valuable insights and critical perspective, which have helped us design a user-friendly and accessible tool beyond purely technical considerations. We also extend our thanks to everyone who supported us throughout the project, despite the repeated delays in production release.

Funding information

The Institut Français de Bioinformatique (IFB) is funded by the Programme d'Investissements d'Avenir (PIA), grant Agence Nationale de la Recherche, number ANR-11-INBS-0013.

This work was supported by the French government grant by the Agence Nationale de la Recherche under France 2030 for structuring research facilities / EQUIPEX+, reference ANR-21-ESRE-0048, and a grant from the Cellule Science Ouverte of the Université Clermont Auvergne (UCA).

The OpenLink project (A gateway between imaging data management tools to apply FAIR principles), an ancestor of Madbot, has been funded by the French Agence Nationale de la Recherche (grant ANR-19-DATA-0011-01) as part of the Flash call "Open Science: research practices and open data".

References

1. Yuan D, Ahamed A, Burgin J, Cummins C, Devraj R, Gueye K, et al. The European Nucleotide Archive in 2023. Nucleic Acids Res. 2024 Jan 5;52(D1):D92–7.

2. European Organization For Nuclear Research, OpenAIRE. Zenodo: Research. Shared. [Internet]. CERN; 2013 [cited 2025 Mar 28]. Available from: https://www.zenodo.org/

3. King G. An Introduction to the Dataverse Network as an Infrastructure for Data Sharing. Sociol Methods Res. 2007 Nov 1;36(2):173–99.

4. Onedata [Internet]. [cited 2025 Mar 3]. Available from: https://onedata.org/#/home

5. iRODS [Internet]. [cited 2025 Mar 3]. Available from: https://irods.org/

6. seek4science.org [Internet]. [cited 2025 Mar 3]. FAIRDOM-SEEK. Available from: https://seek4science.org/

7. elixir-europe/MARS [Internet]. ELIXIR Europe; 2025 [cited 2025 Mar 3]. Available from: https://github.com/elixir-europe/MARS

8. Auffret, Pauline. athENA: FAIR (meta)data management & automatic submission to EBI-ENA [Internet]. Available from: https://gitlab.ifremer.fr/bioinfo/workflows/athena

9. Quiñones M, Liou DT, Shyu C, Kim W, Vujkovic-Cvijin I, Belkaid Y, et al. “METAGENOTE: a simplified web platform for metadata annotation of genomic samples and streamlined submission to NCBI’s sequence read archive”. BMC Bioinformatics. 2020 Sep 3;21(1):378.

10. Jacob D, Ehrenmann F, David R, Tran J, Mirande-Ney C, Chaumeil P. An ecosystem for producing and sharing metadata within the web of FAIR Data. GigaScience. 2025 Jan 6;14:giae111.

11. Celery - Distributed Task Queue — Celery 5.4.0 documentation [Internet]. [cited 2025 Mar 28]. Available from: https://docs.celeryq.dev/en/stable/

12. Redis [Internet]. [cited 2025 Mar 28]. Landing Page - Get Started1. Available from: https://redis.io/lp/get-started1/

13. hashicorp/vault [Internet]. HashiCorp; 2025 [cited 2025 Mar 3]. Available from: https://github.com/hashicorp/vault

14. The Galaxy Community. The Galaxy platform for accessible, reproducible, and collaborative data analyses: 2024 update. Nucleic Acids Res. 2024 Jul 5;52(W1):W83–94.

15. ASGI (Asynchronous Server Gateway Interface) Specification — ASGI 3.0 documentation [Internet]. [cited 2025 Mar 28]. Available from: https://asgi.readthedocs.io/en/latest/specs/main.html

16. Gunicorn - Python WSGI HTTP Server for UNIX [Internet]. [cited 2025 Mar 28]. Available from: https://gunicorn.org/

17. Nuxt [Internet]. [cited 2025 Mar 28]. Nuxt: The Progressive Web Framework. Available from: https://nuxt.com/

18. Vuetify [Internet]. [cited 2025 Mar 28]. Vuetify — A Vue Component Framework. Available from: https://vuetifyjs.com/en/

19. astral-sh/ruff [Internet]. Astral; 2025 [cited 2025 Mar 12]. Available from: https://github.com/astral-sh/ruff

20. Find and fix problems in your JavaScript code - ESLint - Pluggable JavaScript Linter [Internet]. 2025 [cited 2025 Mar 12]. Available from: https://eslint.org/

21. pyupio/safety [Internet]. Safety Cybersecurity (formerly pyup.io); 2025 [cited 2025 Mar 12]. Available from: https://github.com/pyupio/safety

22. PyCQA/bandit [Internet]. Python Code Quality Authority; 2025 [cited 2025 Mar 12]. Available from: https://github.com/PyCQA/bandit

23. Docker: Accelerated Container Application Development [Internet]. 2022 [cited 2025 Mar 28]. Available from: https://www.docker.com/

24. OpenAPI Specification - Version 3.1.0 | Swagger [Internet]. [cited 2025 Mar 6]. Available from: https://swagger.io/specification/

25. API Documentation & Design Tools for Teams | Swagger [Internet]. [cited 2025 Mar 6]. Available from: https://swagger.io/

26. jazzband/django-oauth-toolkit [Internet]. Jazzband; 2025 [cited 2025 Mar 12]. Available from: https://github.com/jazzband/django-oauth-toolkit

27. Penners R. pennersr/django-allauth [Internet]. 2025 [cited 2025 Mar 12]. Available from: https://github.com/pennersr/django-allauth

28. figshare [Internet]. ORCID; 2021 [cited 2025 Mar 28]. From Vision to Value: ORCID's 2022–2025 Strategic Plan. Available from: https://orcid.figshare.com/articles/online_resource/From_Vision_to_Value_ORCID_s_2022_2025_Strategic_Plan/16687207/1

29. LS Login | LifeScience RI [Internet]. [cited 2025 Mar 28]. Available from: https://lifescience-ri.eu/ls-login/

30. JSON Schema [Internet]. [cited 2025 Mar 3]. Available from: https://json-schema.org/

31. Sansone SA, Rocca-Serra P, Gonzalez-Beltran A, Johnson D, ISA Community. Isa Model And Serialization Specifications 1.0. 2016 Oct 28 [cited 2025 Mar 3]; Available from: https://zenodo.org/record/163640

32. Allan C, Burel JM, Moore J, Blackburn C, Linkert M, Loynton S, et al. OMERO: flexible, model-driven data management for experimental biology. Nat Methods. 2012 Mar;9(3):245–53.

33. Bize A, Perréal G, Gramusset A, Predhumeau M, Midoux C, Loux V, et al. DeepOmics, a Digital Environmental Engineering Platform for meta-omics data. In 2020 [cited 2025 Mar 28]. Available from: https://hal.inrae.fr/hal-04493150

34. GitHub [Internet]. [cited 2025 Mar 28]. Build software better, together. Available from: https://github.com

35. GitLab [Internet]. 2023 [cited 2025 Mar 28]. GitLab. Available from: https://gitlab.com/

36. Nextcloud [Internet]. [cited 2025 Mar 28]. Nextcloud - Online collaboration platform. Available from: https://nextcloud.com/fr/

37. PRIDE database at 20 years: 2025 update | Nucleic Acids Research | Oxford Academic [Internet]. [cited 2025 Mar 28]. Available from: https://academic.oup.com/nar/article/53/D1/D543/7874848?login=false

38. Sansone SA, McQuilton P, Rocca-Serra P, Gonzalez-Beltran A, Izzo M, Lister AL, et al. FAIRsharing as a community approach to standards, repositories and policies. Nat Biotechnol. 2019 Apr;37(4):358–67.

39. Research Organization Registry (ROR) [Internet]. [cited 2025 Mar 28]. Research Organization Registry (ROR). Available from: https://ror.org/

40. DOI foundation [Internet]. [cited 2025 Mar 28]. Available from: https://www.doi.org/the-foundation/about-us/

41. PubMed [Internet]. [cited 2025 Mar 28]. PubMed. Available from: https://pubmed.ncbi.nlm.nih.gov/

42. Home - Taxonomy - NCBI [Internet]. [cited 2025 Mar 28]. Available from: https://www.ncbi.nlm.nih.gov/taxonomy

43. Inc B. Lab Management Software | Laboratory System | Labguru [Internet]. [cited 2025 Mar 28]. Available from: https://www.labguru.com

44. eLabFTW - Open Source Laboratory Notebook [Internet]. [cited 2025 Mar 28]. Available from: https://www.elabftw.net

45. Seafile - Open Source File Sync and Share Software [Internet]. [cited 2025 Mar 28]. Available from: https://www.seafile.com/en/home/

46. OPIDoR/DMPOPIDoR [Internet]. OPIDoR; 2024 [cited 2025 Mar 28]. Available from: https://github.com/OPIDoR/DMPOPIDoR

47. GitHub [Internet]. [cited 2025 Mar 28]. Data Stewardship Wizard. Available from: https://github.com/ds-wizard

48. Packaging research artefacts with RO-Crate - Silvio Peroni, Stian Soiland-Reyes, Peter Sefton, Mercè Crosas, Leyla Jael Castro, Frederik Coppens, José M. Fernández, Daniel Garijo, Björn Grüning, Marco La Rosa, Simone Leo, Eoghan Ó Carragáin, Marc Portier, Ana Trisovic, RO-Crate Community, Paul Groth, Carole Goble, 2022 [Internet]. [cited 2025 Mar 28]. Available from: https://journals.sagepub.com/doi/10.3233/DS-210053

49. json-ld/json-ld.org [Internet]. JSON-LD Community; 2025 [cited 2025 Mar 28]. Available from: https://github.com/json-ld/json-ld.org

50. Amazon Web Services, Inc. [Internet]. [cited 2025 Mar 28]. Services et produits de cloud Amazon | AWS. Available from: https://aws.amazon.com/fr/

Session 8: Statistics, Machine Learning, and AI for Biology and Health

Benchmarking Data Leakage on Link Prediction in Biomedical Knowledge Graph Embeddings

Galadriel BRIERE[1], Thomas STOSSKOPF[1], Benjamin LOIRE[1] and Anaïs BAUDOT[1]

1 Aix Marseille Univ, INSERM, MMG, Marseille, France

Corresponding author: marie-galadriel.briere@univ-amu.fr

Keywords

Knowledge Graph Embedding, Data Leakage, Link Prediction, Drug Repurposing, Rare Diseases

Abstract

In recent years, Knowledge Graphs (KGs) have gained significant attention for their ability to organize complex biomedical knowledge into entities and relationships. Knowledge Graph Embedding (KGE) models facilitate efficient exploration of KGs by learning compact data representations. These models are increasingly applied to biomedical KGs for link prediction, for instance to uncover new therapeutic uses for existing drugs. While numerous KGE models have been developed and benchmarked for link prediction, existing evaluations often overlook the critical issue of data leakage. Data leakage leads the model to learn patterns it would not encounter when deployed in real-world settings, artificially inflating performance metrics and compromising the overall validity of benchmark results. In machine learning, data leakage can arise when (1) there is inadequate separation between training and test sets, (2) the model leverages illegitimate features, or (3) the test set does not accurately reflect real-world inference scenarios. In this study, we implement a systematic procedure to control train-test separation for KGE-based link prediction and demonstrate its impact on models' performance. In addition, through permutation experiments, we investigate the potential use of node degree as an illegitimate predictive feature, finding no evidence of such leveraging. Finally, by evaluating KGE models on a curated dataset of rare disease drug indications, we demonstrate that performance metrics achieved on real-world drug repurposing tasks are substantially worse than those obtained on drug-disease indications sampled from the KG.

RITHMS : An advanced stochastic framework for the simulation of transgenerational hologenomic data

Solène PETY[1,2], Ingrid DAVID[3], Andrea RAU[1], and Mahendra MARIADASSOU[2]

1 Université Paris-Saclay, INRAE, GABI, 78350, Jouy-en-Josas, France

2 Université Paris-Saclay, INRAE, MaIAGE, 78350, Jouy-en-Josas, France

3 Université de Toulouse, INRAE, ENVT, GenPhySE, 31326, Castanet-Tolosan, France

Corresponding author: *solene.pety@inrae.fr*

Keywords

Holobiont, genotypes, microbiota data, simulation framework

Abstract

A holobiont is made up of a host organism together with its microbiota. In the context of animal breeding, the holobiont can be viewed as the single unit upon which selection operates. Therefore, integrating microbiota data into genomic prediction models may be a promising approach to improve predictions of phenotypic and genetic values. Nevertheless, there is a paucity of hologenomic transgenerational data to address this hypothesis, and thus to fill this gap, we propose a new simulation framework. Our approach, an R Implementation of a Transgenerational Hologenomic Model-based Simulator (RITHMS) is an open-source package, builds upon simulated transgenerational genotypes from the MoBPS package and incorporates distinctive characteristics of the microbiota, notably vertical and horizontal transmission as well as modulation due to the environment and host genetics. In addition, RITHMS can account for a variety of selection strategies and is adaptable to different genetic architectures. We simulated transgenerational hologenomic data using RITHMS under a wide variety of scenarios, varying heritability, microbiability, and microbiota heritability. We found that simulated data accurately preserved key characteristics across generations, notably microbial diversity metrics, exhibited the expected behavior in terms and correlation between taxa and of modulation of vertical and horizontal transmission, response to environmental effects and the evolution of phenotypic values depending on selection strategy. Our results support the relevance of our simulation framework and illustrate its possible use for building a selection index balancing genetic gain and microbial diversity. RITHMS is an advanced, flexible tool for generating transgenerational hologenomic data that incorporate the complex interplay between genetics, microbiota and environment.

Variable selection in transcriptomics data using knockoffs in a classification framework

Julie CARTIER[1,2,3], Chloé-Agathe AZENCOTT[1,2,3], Adeline FERMANIAN[4], Johanna LAGOAS[5] and Florian MASSIP[1,2,3]

1 Centre for Computational Biology (CBIO), Mines Paris, PSL University, 60 bd Saint-Michel, 75272, Paris, France

2 Institut Curie, PSL University, 11 rue Pierre et Marie Curie, 75005, Paris, France

3 U1331, INSERM, 11 rue Pierre et Marie Curie, 75005, Paris, France

4 LOPF, Califrais' Machine Learning Lab, Paris, France

5 AgroParisTech, Paris-Saclay University, 22 place de l'Agronomie, 91 123, Palaiseau, France

Corresponding author: julie.cartier@minesparis.psl.eu

Keywords

Variable selection, Knockoffs, Transcriptomic

Abstract

The emergence of new sequencing technologies has facilitated the acquisition of large amounts of biological data, which has proven to be a useful tool for better understanding biological systems. One way to take advantage of the potential of sequencing data is to use them to identify the relationship between biological units (e.g. genes) and phenotypical characteristics (e.g. disease outcomes). This question, formulated as a variable selection problem, remains difficult because of the size of the data ($n \ll p$) and their correlation structure. To address these challenges, we studied the applicability of the knockoff (KO) procedure focusing on transcriptomic data in a classification setting. Introduced by Candès et al. in 2015, the KO variable selection procedure has shown promising results on real biological data. This method seeks to identify the truly important predictors by overcoming the correlation structure between variables while controlling the false discovery rate even in high dimensional settings. We conducted an extensive simulation study using real data to evaluate the relevance of recent methods in the context of high-dimensional classification. We also analyzed the benefits of a KO aggregation scheme to mitigate the effect of stochasticity, which is intrinsic to the KO procedure. In addition, we studied the stability of the KO framework as a measure of the reliability of variable selection. Finally, we applied the KO framework to real transcriptomic data.

Evaluating deep learning models for plant protein function prediction

Minh Ngoc VU[1], Hoang Ha NGUYEN[2], Antoine TOFFANO[3] and Pierre LARMANDE[3,4]

1 Université Evry Paris-Saclay, 91000, Evry-Courcouronnes, France

2 ICT Department, University of Science and Technology of Hanoi, 100000, Hanoi, Vietnam

3 LIRMM, Univ. Montpellier, 34000, Montpellier, France

4 DIADE, IRD, CIRAD, Univ. Montpellier, 34000, Montpellier, France

Corresponding author: pierre.larmande@ird.fr

Keywords

Protein function prediction, Gene Ontology, deep learning

Abstract

Predicting the functions of proteins remains a critical yet challenging task in computational biology. Advances in high-throughput sequencing, the expansion of protein databases, and the continuous development of artificial intelligence have led to the emergence of many computational methods dedicated to protein function prediction. In this study, we evaluated the performance of four state-of-the-art models — DeepGOPlus, DeepGraphGO, DeepGOZero, and DeepGOSE — using experimentally annotated proteins from the UniProt-KB/Swiss-Prot database. We also trained and tested these models on species-specific datasets from Arabidopsis thaliana and Oryza sativa to investigate their potential and applicability in plant protein studies. Our results showed that DeepGOPlus consistently achieved the best evaluation scores across all datasets. DeepGOSE and DeepGOZero performed comparably and only marginally outperformed DeepGraphGO in certain training attempts. Further analysis revealed that dataset stratification into training, validation, and testing sets introduced variations in Gene Ontology annotation specificity, which may have influenced model performance.

jsPCA enables fast, interpretable and parameter-free domain identification in 3D spatial transcriptomics data

Ines ASSALI[1], Paul ESCANDE[2] and Paul VILLOUTREIX[1]

1 Aix-Marseille Université, MMG, Inserm U1251, Turing Centre for Living systems, Marseille, France

2 Institut de Mathématiques de Toulouse; UMR 5219, Univ. de Toulouse, CNRS ; UPS, F-31062 Toulouse Cedex 9, France

Corresponding author: ines.assali@univ-amu.fr, paul.villoutreix@univ-amu.fr

Keywords

Spatial transcriptomics, Spatial statistics, Machine learning

Abstract

Spatial Transcriptomics (ST) uncovers gene expression patterns within the structured spatial layout of tissues, a level of detail absent in single-cell transcriptomics analysis, which enhances our comprehension of cell-environment interactions. Accurate spatial information is critical for clustering cell domains and for a better understanding of their functional connections in intricate biological tissues. In this study, we propose a novel approach, joint spatial principal component analysis (jsPCA), to efficiently reveal complex gene expression profiles while preserving the spatial context of tissues in multi-slices or multi-samples ST. Our approach consists in identifying the principal components (PCs) that best maximize the product of spatial autocorrelation (Moran's Index) and transcriptomic covariance, reflecting both the structure of genetic expression and its spatial distribution. By combining dimensionality reduction and emphasis on spatial correlations, jsPCA refines the ability to detect spatial gene expression patterns and variations, thereby improving the outcome of domain clustering. We take advantage of sparse matrices to improve scalability, which makes it ideally adapted to the analysis of large-scale ST datasets. The interpretability of jsPCA arises from its linear structure, which provides a clear understanding of the impact of each variable on the clustering results, in contrast to current more complex approaches based on Graph Neural Networks. jsPCA handles multi-slice or multi-sample analysis. Spatial domains are obtained by Gaussian mixture clustering in this joint space. We evaluated our approach using the Visium 10x dataset of human dorsolateral prefrontal cortex (DLPFC), featured in numerous benchmarks. Our approach demonstrated robust performance, comparable or better to various state-of-the-art methods, while being fast, interpretable and parameter free.

Leveraging multi-omics integration to uncover childhood trauma-related mechanisms in bipolar disorder.

Margot DEROUIN[1], Amazigh MOKHTARI[1], El Cherif IBRAHIM[2], Pierre-Eric LUTZ[3,4], Raoul BELZEAUX[2,5], Cynthia MARIE-CLAIRE[6], Frank BELLIVIER[6,7,8], Bruno ETAIN[6,9], Cathy PHILIPPE1[0], Andrée DELAHAYE-DURIEZ[1,11,12]

1 Université Paris Cité, Inserm, NeuroDiderot, UMR-1141, 75019, Paris, France.

2 Aix-Marseille Univ, CNRS, INT, Inst Neurosci Timone, 13005, Marseille, France.

3 Centre National de la Recherche Scientifique, Université de Strasbourg, Institut des Neurosciences Cellulaires et Intégratives UPR 3212, F-67000, Strasbourg, France;

4 Douglas Mental Health University Institute, McGill University, QC, H4H 1R3, Montréal, Canada.

5 Département de psychiatrie, CHU de Montpellier, Montpellier, France.

6 Université Paris Cité, INSERM UMR-S 1144, Optimisation thérapeutique en neuropsychopharmacologie, OTeN, F-75006, Paris, France.

7 Fondation FondaMental, Créteil, France;

8 AP-HP, Groupe Hospitalo-Universitaire AP-HP Nord, DMU Neurosciences, Hôpital Fernand Widal, Département de Psychiatrie et de Médecine Addictologique, Paris, France.

9 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.

10 Université Paris-Saclay, CEA, CNRS, Neurospin, Baobab UMR 9027, Gif-sur-Yvette, France.

11 Unité fonctionnelle de médecine génomique et génétique clinique, Hôpital Jean Verdier, Assistance Publique des Hôpitaux de Paris, F-93140, Bondy, France;

12 Université Sorbonne Paris Nord, F-93000, Bobigny, France

Corresponding Author: margot.derouin@inserm.fr

Keywords

Multi-omics Integration, Sample size, Bipolar Disorder, DNA methylation, Transcriptomic

Abstract

Background : Childhood trauma, including abuse or neglect, has profound effects on mental health, increasing susceptibility to psychiatric disorders. Bipolar disorder, marked by extreme mood swings encompassing manic and depressive episodes, disrupts daily functioning. Despite the growing interest in molecular psychiatry, the etiology of bipolar disorder remains unclear, with no established blood biomarkers [1]. This gap of knowledge is partially due to the complexity and heterogeneity of the

disorder. Additionally, environmental factors, particularly early-life trauma, are suspected to play a significant role in the onset and progression of bipolar disorder [2,3,4].

Recent advances in Next-Generation Sequencing (NGS) have generated extensive genomic data, yet the integration of multi-omic data with advanced machine learning techniques remains underutilized in psychiatric research[5]. As seen in cancer research, the application of multi-omics approaches that combine genetic, transcriptomic, and epigenomic data with machine learning holds potential for advancing our understanding of psychiatric disorders.

Material and Methods : This study aims to determine the minimum sample size required to accurately predict trauma exposure and identify potential biomarkers of childhood trauma in peripheral blood samples from bipolar patients. We utilized transcriptomics (RNA-seq), and epigenomics (miRNA-seq and DNA methylation) datasets from a cohort of bipolar disorder (n = 274) patients, all of whom were assessed using the Childhood Trauma Questionnaire (CTQ). After quality control and preprocessing the final dataset included 200 individuals with DNAm data, 122 individuals with mRNA and miRNA data, and 102 individuals with data from all three omics modalities.

We derived train/test subsets by gradually increasing the sample size in the training set. Using an advanced joint reduction dimension method, named Regularized Generalized Canonical Correlation Analysis (RGCCA) [6], we evaluated the prediction error rates as a function of sample size in the training set.

Results and Discussion : The analysis revealed that N80% = 81 individuals in the training set (i. e. 80% train-test split), for at least 2 modalities over 3 and from 3 components per block, achieved the best prediction performances. However, almost no feature survived the multiple testing procedure when assessing model stability, suggesting that further investigations are needed to obtain a biologically interpretable sparse model.

References

[1] Legrand A, Iftimovici A, Khayachi A, Chaumette B. Epigenetics in bipolar disorder: a critical review of the literature.Psychiatr Genet. 2021 Feb 1;31(1):1-12.

[2] Aas M, Henry C, Andreassen OA, Bellivier F, Melle I, Etain B. The role of childhood trauma in bipolar disorders. Int J Bipolar Disord. 2016 Dec;4(1):2.

[3] Agnew-Blais J, Danese A. Childhood maltreatment and unfavourable clinical outcomes in bipolar disorder: a systematic review and meta-analysis. Lancet Psychiatry. 2016 Apr;3(4):342-9.

[4] Palmier-Claus JE, Berry K, Bucci S, Mansell W, Varese F. Relationship between childhood adversity and bipolar affective disorder: systematic review and meta-analysis. Br J Psychiatry. 2016 Dec;209(6):454-459.

[5] Mokhtari, A., Porte, B., Belzeaux, R., Etain, B., Ibrahim, E.C., Marie-Claire, C., Lutz, P.E., Delahaye-Duriez, A., 2022. The molecular pathophysiology of mood disorders: From the analysis of single molecular layers to multi-omic integration. Prog. Neuro- Psychopharmacology Biol. Psychiatry 116.

[6] Tenenhaus, A., & Tenenhaus, M. (2011). Regularized Generalized Canonical Correlation Analysis. Psychometrika, 76(2), 257–284.

Models for protein domain embedding

Louison SILLY[1,2], Guy PERRIERE[1] and Philippe ORTET[2]

1 Laboratoire de Biométrie et de Biologie évolutive, Université Claude Bernard Lyon 1, 43 bd. du 11 novembre 1918, 69622, Villeurbanne, France

2 Aix-Marseille Université, CNRS, CEA, BIAM, UMR7265 Institut de Biosciences and Biotechnologies d'Aix-Marseille, Cadarache research centre, F-13115 Saint-Paul-lez-Durance, France

Corresponding author: philippe.ortet@cea.fr

Keywords

Deep Learning, protein embedding, protein annotation, protein domains

Abstract

Protein embedding consist of producing a mathematical representation of a protein based on data such as sequence or structure. Protein embedding is widely use thanks to the last advances in artificial intelligence that allow embedding to resume proteins information (such as function, biochemical properties, ...). One of the setback to protein embedding is the dimensionality of the latest, often very large, leading to difficulties to manipulate them efficiently (the so called Curse of Dimensionality). We present here our work on protein embedding based on protein domain architecture. A protein having less domains than amino acids we hope to produced embedding of lower dimensionality that would be easier to use. We trained two models based on the Bert architecture on different training datasets using the mask language modeling objective. Our training datasets were obtained by annotating Uniprot (Trembl + SwissProt) and BFD proteins using PFAM domains and Low Complexity Region. Our models show good performances on some training sets and seems to be able to learn a good protein representation from their domains architecture.

Exhaustive Identification of Pleiotropic Loci for Serum Leptin Levels in the NHGRI-EBI Genome-Wide Association Catalog

Anthony HAIDAMOUS[1], David MEYRE[1,2], Sébastien HERGALANT[1]

1 INSERM U1256, NGERE, Université de Lorraine, 54000, Nancy, France

2 Service de Biochimie et Biologie Moléculaire, CHRU de Nancy, 54000, Nancy, France

Corresponding author: sebastien.hergalant@univ-lorraine.fr

Keywords

Leptin, Genome-wide associations studies, Gene pleiotropy, Variant effect, Protein-protein interaction modules.

Abstract

Leptin is an adipokine that regulates energy expenditure and calory intake by acting on the hypothalamic leptin-melanocortin pathway. It has known effects on obesity, inflammation, and neurodevelopment, hence carrying high pleiotropic potential. Using the Genome-Wide Association Study (GWAS) Catalog, we identified the single nucleotide polymorphisms (SNPs) associated with fluctuations of leptin in the blood, which we expanded into wider genetic blocks, and explored their associations and effect sizes with leptin levels. We then investigated the genetic pleiotropy linked with these genomic regions.

Starting from 35 GWAS studies on leptin levels with a minimum discovery sample size > 1000 individuals, we selected 25 SNPs reaching genome-wide significance ($p < 5.10\text{-}8$). This led to the aggregation of 15 genetic blocks with SNPs in high linkage disequilibrium with the leptin SNPs. The blocks were also associated with 574 pleiotropic phenotypes, which were then grouped into 22 categories, including the enriched “other adipokines”, “obesity-related”, “inflammation-related”, “cancer-related”, “body fat”, “type-2-diabetes-related”, and “addiction-related” cross-traits. The list of genes overlapping with the genetic blocks was used to map a protein-protein interaction network surrounding leptin with which we identified functional modules enriched in ontological terms such as the leptin-melanocortin pathway, embryogenesis, immunity and transcription regulation.

This study extends the genetic architecture behind leptin levels to its wider roles in human physiology by deciphering molecular pathways and gene modules implicated in its end effects.

Introduction

Obesity is a pandemic that affects more than 38% of the global population (World Obesity Atlas 2023 report (1)). It is defined as an abnormal accumulation of body fat and is associated with complexities and morbid diseases, such as type 2 diabetes (T2D) (2,3), cardiovascular diseases (4) and cancers (5–8). A key player in the pathophysiology of obesity is leptin, a small adipokine produced from the *LEP* gene (9) almost exclusively by the adipose tissue, and released in the serum after energy intake (10,11). Leptin acts primarily in the hypothalamus to activate neuroendocrine pathways that will insure its main functions of satiety and lipid metabolism control (12). In order to understand the factors that affect serum leptin levels and decipher the pleiotropic effects surrounding its physiological expression, we developed a method that exhaustively identifies all the leptin-associated single nucleotide polymorphisms (SNPs) reported by Genome-wide associations studies (GWAS) and infers genotype-phenotype correlations (13) between leptin and a number of related traits, such as obesity (14,15), T2D and insulin levels (15,16), cardiovascular diseases (17), and inflammation (3). Among these associations, some involve pleiotropic genes, *i.e.* influencing multiple functions and/or multiple phenotypes. Current GWAS and pleiotropy analyses can identify associations among traits and variants but lack the resolution to untangle the links between groups of genetic variants and causal genes as well as biological pathways. The original approach presented in this study calculates genetic blocks of linkage disequilibrium SNPs aggregated from the GWAS leptin loci and extracts cross-phenotypes associations, effect sizes and directions as compared to leptin. It then identifies protein-coding genes within these genomic regions and maps out their pleiotropic potential using a functional enrichment method based on protein-protein interactions, method that also infers new genes associated with leptin while excluding non-interacting genes.

Methods

Genome-wide association studies catalog. The NHGRI-EBI Catalog of Published Genome-Wide Association Studies (GWAS Catalog) (18), up to date from March 11th, 2024, and annotated for the GRCh38/hg38 human reference genome, was used extensively during this study.

R and its packages. GWAS Catalog data were downloaded for local use, handled, and manipulated using R4.3.2 (19). Key packages such as ggplot2 (used for graphs and data visualization), dplyr (used for data cleaning), karyotploteR (20) (used for blocks and genes visualization on chromosomes), and LDlinkR (21) (used for linkage disequilibrium analysis) were used.

LDlink. Within the LDlink suite (https://ldlink.nih.gov/?tab=home, February 7th, 2024) (22), LDproxy is a tool allowing for an exploration of proxy variants associated with an input query SNP. We used GWAS

leptin-associated SNPs as query/lead SNPs (lSNPs) for the calculations, with the 1000 genome project data as reference. Leptin lSNPs were defined as reaching genome-wide significance ($p < 5.10\text{-}8$) in a study using a minimum discovery sample size > 1000 individuals. LDproxy outputs report variants associated to each lSNP ranked by their LD correlation value (r^2) in their respective populations, encoded by ethnicity (EUR: European, AFR: African, AMR: Ad-mixed Americans, EAS: East-Asians, SAS: South Asians) according to the discovery cohorts of the lSNP.

Genetic blocks, variants, and cross-traits. Using the lSNP ($r^2 = 1$), proxy SNPs (pSNPs) ($r^2 \geq 0.9$), and non-independent SNPs ($r^2 \geq 0.1$), genetic blocks of LD were aggregated. If multiple lSNPs were mapped to the same block, we retained a main lead variant as the one with the smallest association p-value. Block ethnicities were assigned based on combined ethnicities for the lead SNPs. Using every lSNPs + pSNPs, we compiled a table of cross-traits (phenotypes/traits/diseases also associated with the variants, but not leptin, $p < 5e\text{-}8$), which was hierarchized into general traits used to form categories of traits.

Fold-enrichment analyses were conducted for traits and categories, giving insight on their representation and distribution within the leptin blocks as compared to what is expected from the background. Enrichments were calculated following the relation: frequency of the observed trait / expected frequency, where the observed frequency represents the number of reported occurrences for the trait in the leptin selection among the total number of reported leptin cross-trait occurrences, and the expected frequency represents the overall significant occurrences for the same trait among the total number of significant associations in the complete GWAS Catalog ($p < 5e\text{-}8$).

$$Fold\ Enrichment\ (e) = \frac{\frac{Occurence\ of\ trait\ in\ leptin\ pSNPs}{Occurence\ of\ trait\ in\ GWAS\ catalog}}{\frac{Total\ number\ of\ cross-traits}{Total\ number\ of\ traits\ in\ GWAS\ catalog}}$$

Two-way Fisher's exact test was used to evaluate the enrichment statistical significance ($p < 0.05$).

Gene-based functional annotations and pleiotropic networks. Using the list of protein coding genes overlapping with the genetic blocks of LD and reported as "mapped genes" with each lSNPs, we formed a network of significantly enriched protein-protein interactions (PPI, enrichment p-value 1e-05) with STRING-db (https://string-db.org/), a repository of functional and physical PPI data (23,24). The generated network was then augmented with first-shell interactors (n = same number as input genes) with a confidence score of 0.5, and clustered using the k-means algorithm for functional module identification. Each cluster was subject to functional annotations using the databases found in EnrichR (25–27) for pathway (Reactome, KEGG, BioPlanet and BioCarta), ontology (GO biological process, cellular component,

and molecular function), and disease (PhenGenI) enrichment. Each cluster was then isolated and augmented with 10 additional first-shell interactors for further characterization (workflow in **Figure 1**).

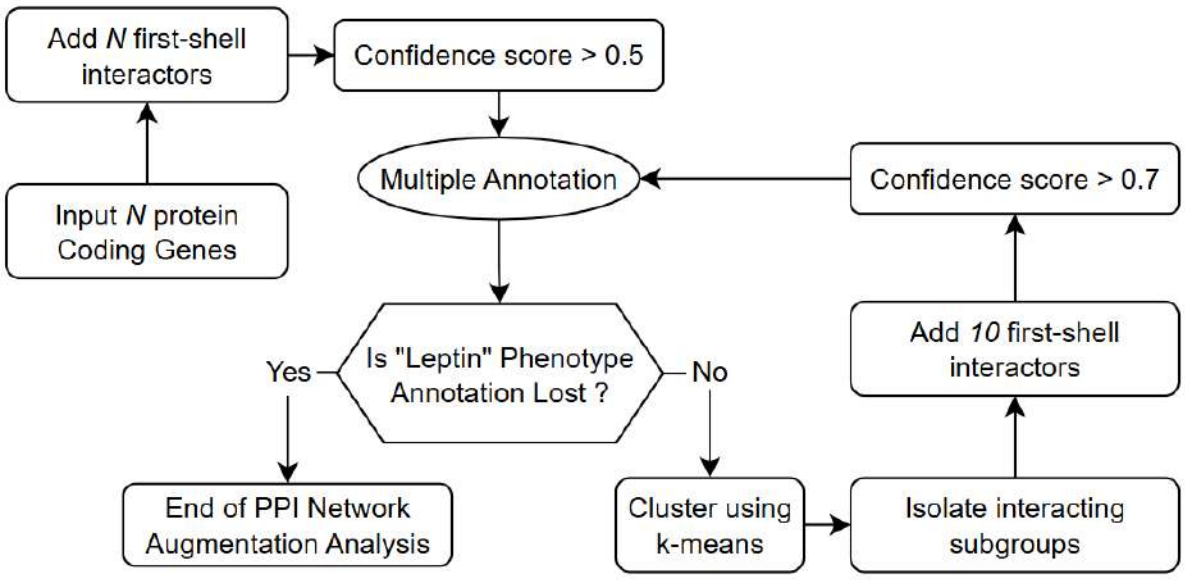


Figure 1. Flowchart describing the algorithm used for constructing and annotating the pleiotropic network around Leptin.

Results

Leptin genetic pleiotropy. Upon selecting significant SNPs mapped to "leptin measurement" and "BMI-adjusted leptin measurement" in the GWAS Catalog (discovery cohort > 1000 individuals; p-value < 5e-8), we obtained a list of 25 lead leptin SNPs (lSNPs) (**Figure 2A**). The list was then used for the LD analysis, which resulted in 343 proxy SNPs (pSNPs; $r^2 \geq 0.9$) and 3714 non-independent SNP ($r^2 \geq 0.1$). All variants were aggregated under their lSNPs into 15 genetic blocks of highly correlated structure (**Figure 2B-C**). Main characteristics for the genetic block constitution is summarized in **Table 1**.

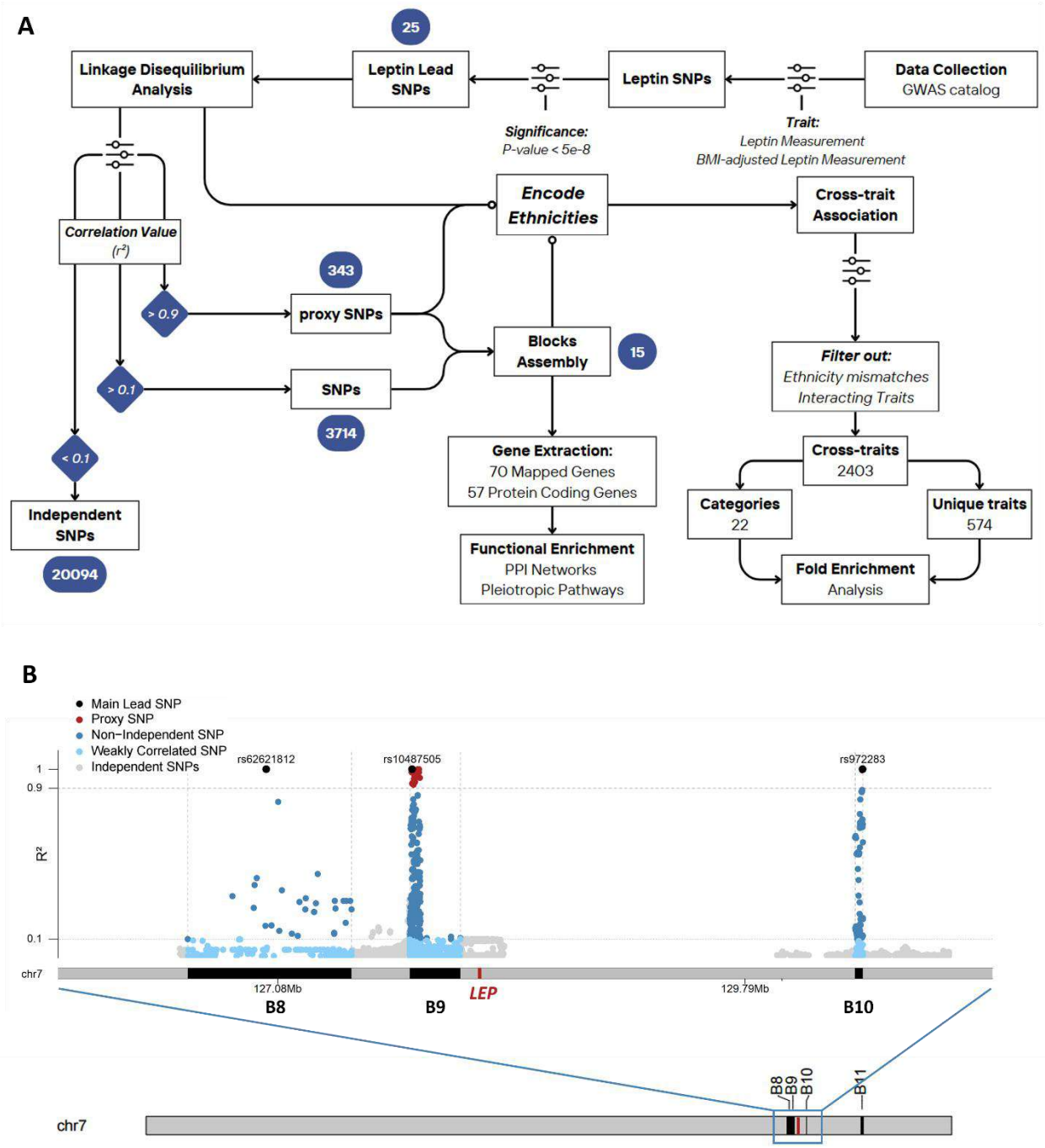


Figure 2. Workflow and Genetic Blocks. (A) Flowchart illustrating the step-by-step extraction of adiponectin SNPs and cross-trait associations, **(B)** Close-up on the structure and LD SNPs of the blocks surrounding the LEP gene.

Block	Genomic Position	Block span (kb)	Number of lSNPs	Number of pSNPs	Number of SNPs	Number of cross-traits	Number of overlapping genes	Ethnicities	Mapped genes
B1	chr1:243096987-243450323	353	2	16	254	0	2	AFR	SDCCAG8, CEP170
B2	chr2:27386799-28205581	819	2	3	525	1481	16	EUR, AFR, EAS, AMR	SNX17, ZNF513, PPM1G, NRBP1, KRTCAP3, FNDC4, IFT172, ZNF512, CCDC121, GPN1, SUPT7L, GCKR, MRPL33, RBKS, SLC4A1AP, BABAM2
B3	chr2:60081471-60559270	478	1	0	100	0	1	AFR	BCL11A
B4	chr2:142550181-143496235	946	1	10	112	0	2	AFR	KYNU, ARHGAP15
B5	chr2:165057377-165908940	852	1	11	356	295	5	EUR, AFR, EAS, AMR	TTC21B, SCN2A, SCN3A, CSRNP3, GALNT3
B6	chr3:156791268-156929068	138	2	16	52	138	1	EUR, AFR, EAS, AMR	LEKR1
B7	chr5:174411013-175372256	961	1	0	37	0	3	AFR	DRD1, SFXN1, MSX2
B8	chr7:126560442-127509095	949	1	0	27	0	7	EUR, AFR, EAS, AMR	SND1, ZNF800, GRM8, GCC1, ARF5, FSCN3, PAX4
B9	chr7:127408546-128138453	730	5	12	137	6	10	EUR	SND1, RBM28, ZNF800, GRM8, GCC1, ARF5, FSCN3, PAX4, LRRC4, **LEP**
B10	chr7:130422934-130468190	45	1	1	111	2	1	EUR	CEP41
B11	chr7:141144363-141714708	570	2	42	297	3	5	AFR	PRSS37, AGK, TMEM178B, DENND11, WEE2
B12	chr12:33789141-34775707	987	1	14	625	0	1	AFR	ALG10
B13	chr14:79416899-80182942	766	1	0	77	0	1	AFR	NRXN3
B14	chr16:53773114-53848561	75	3	85	170	478	1	EUR	FTO
B15	chr22:17276066-17770695	495	1	1	7	0	5	AFR	SLC25A18, ATP6V1E1, BCL2L13, BID, CECR2

Table 1. Leptin genetic bloc characteristics. See Methods for ethnicity encodings.

Cross-trait analysis. Using the list of lSNPs + pSNPs, we associated 574 pleiotropic phenotypes (or cross-traits) within these blocks, which we categorized into 22 main items. The cross-trait occurrences were checked for over-representation against background (the entire significant GWAS Catalog associations), which uncovered a number of significantly enriched categories containing traits related to “diet or taste” (fold-enrichment e = 4.4), “other adipokines” (e = 3.5), “addiction” (e = 2.4), “body fat” (e = 2.3), “inflammation” (e = 2), “lipids” (e = 1.7) and “cancer-related” (e = 1.7), as well as some depleted categories related to “brain” (e = 0.1), “blood pressure” (e = 0.2), “sleep” (e = 0.3) and “age” (e = 0.3) (**Figure 3A**). This summary, however, is not representative of each unique trait, as the specific obesity trait, belonging to the “anthropometric measurement” category, is highly over-represented (e = 15) among other under-represented traits (**Figure 3B**). Next, we summarized cross-traits with their respective associated-SNP beta coefficients to study the direction of effect of each phenotype as compared to leptin (**Figure 3C**). Consistent monodirectional effects with leptin were determined with body mass index (mean effect with leptin (me) = +0.063), body fat percentage (me = +0.342), BMI-adjusted hip circumference (me = +0.024),

bone mineral density (me = +0.025), and obesity (me = +0.277). Opposite directions, however, were found with hyperlipidemia (me = 0.253), insulin-like growth factor-binding protein 1 (me = -0.134), insulin sensitivity measurement (me = -0.200), and the other adipokines, adiponectin (me = -0.107) and leptin receptor measurements (me = -0.114). Finally, some traits present mixed effects with leptin levels, such as type-2 diabetes, addiction-related traits (intakes of alcohol and coffee, smoking behavior, substance use), high density lipoprotein cholesterol, and some taste preferences such as sweetened drinks and sugar intake. The distribution of the cross-trait across leptin genetic blocks was heterogeneous (**Table 1**).

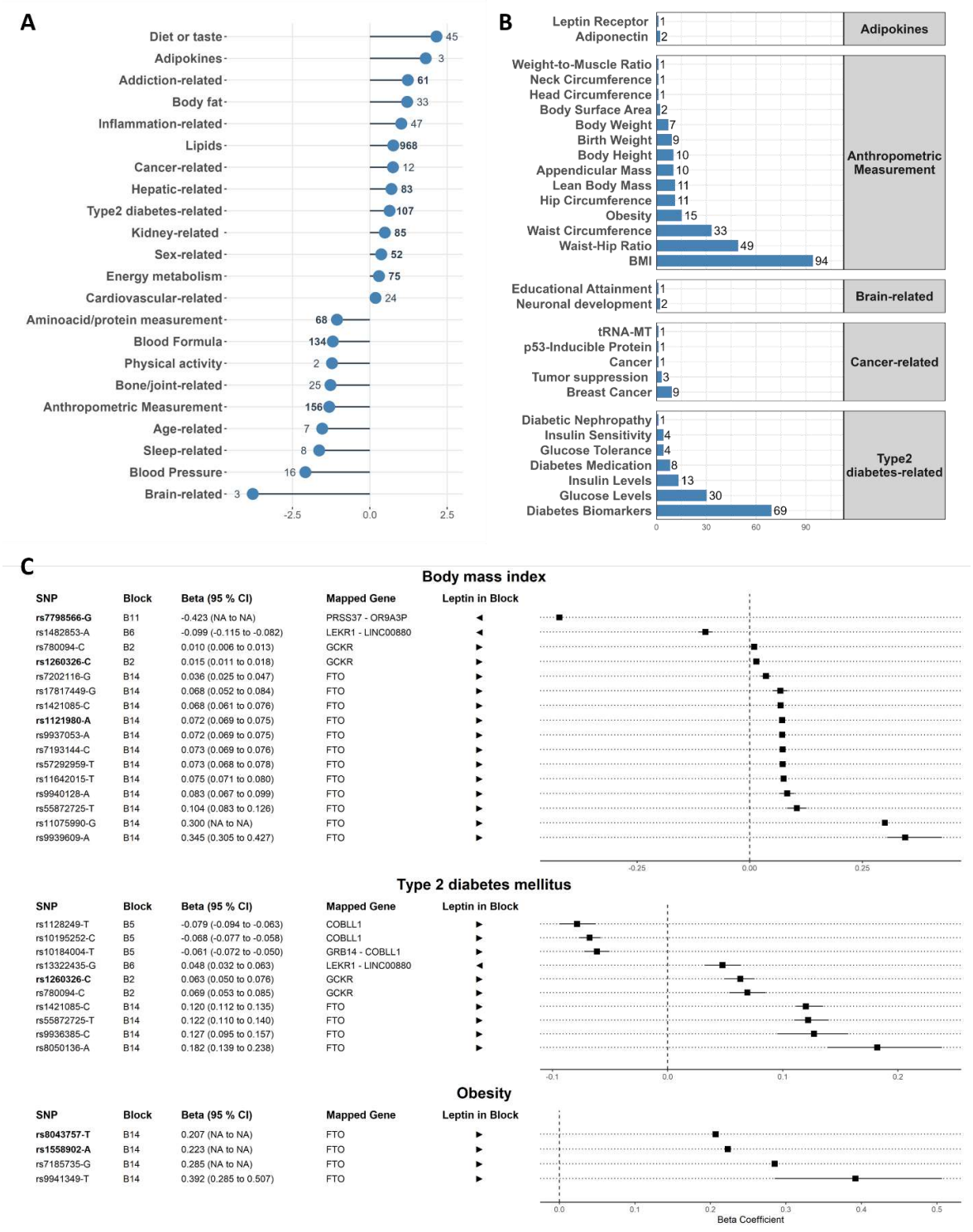


Figure 3. Cross-traits associated with leptin levels. (A) Normalized fold-enrichment values for the 22 categories of traits associated with leptin measurement and BMI-adjusted leptin measurement. **(B)** Raw counts for "adipokines", "anthropometric measurement", "brain-related", "cancer-related", "body fat" and "type-2-diabetes-related" cross-trait categories. **(C)** Forest plots showing the direction of effect of the traits as compared to leptin levels in each genetic block. Beta coefficients represent the increase or decrease per unit in the outcome for each significantly associated trait (p<5e-08).

Gene-based pleiotropic network. Concomitantly to the cross-trait analysis, we conducted a protein network reconstruction around leptin. We first extracted 57 protein-coding genes overlapping the genetic blocks of leptin (**Table 1,** 4 genes were associated with consecutive blocks). These input genes were used in STRING-db to map out the functional protein modules involved in leptin molecular functions (**Figure 1**, see Methods). A preliminary functional annotation of the network was done using the Phenotype-Genotype Integrator (PheGenI) library, to explore the diseases and traits associated with the input genes (**Figure 4A**). The analysis confirmed a highly significant association with the leptin phenotype (e = 19), and revealed other enriched associations with phenotypes including type-2 diabetes, body-mass index, obesity, abdominal fat and metabolic syndrome X. This initial network was augmented with 57 first-shell interactors at a confidence score of 0.5, and clustered using the k-means algorithm. This network expansion identified newly enriched or depleted, as well as gain/loss of PhenGenI terms (**Figure 4B**), with a noticeable decrease in the leptin phenotype (e = 9.5), and increases in metabolic syndrome X, type-2 diabetes, and glucose tolerance test (term gain).

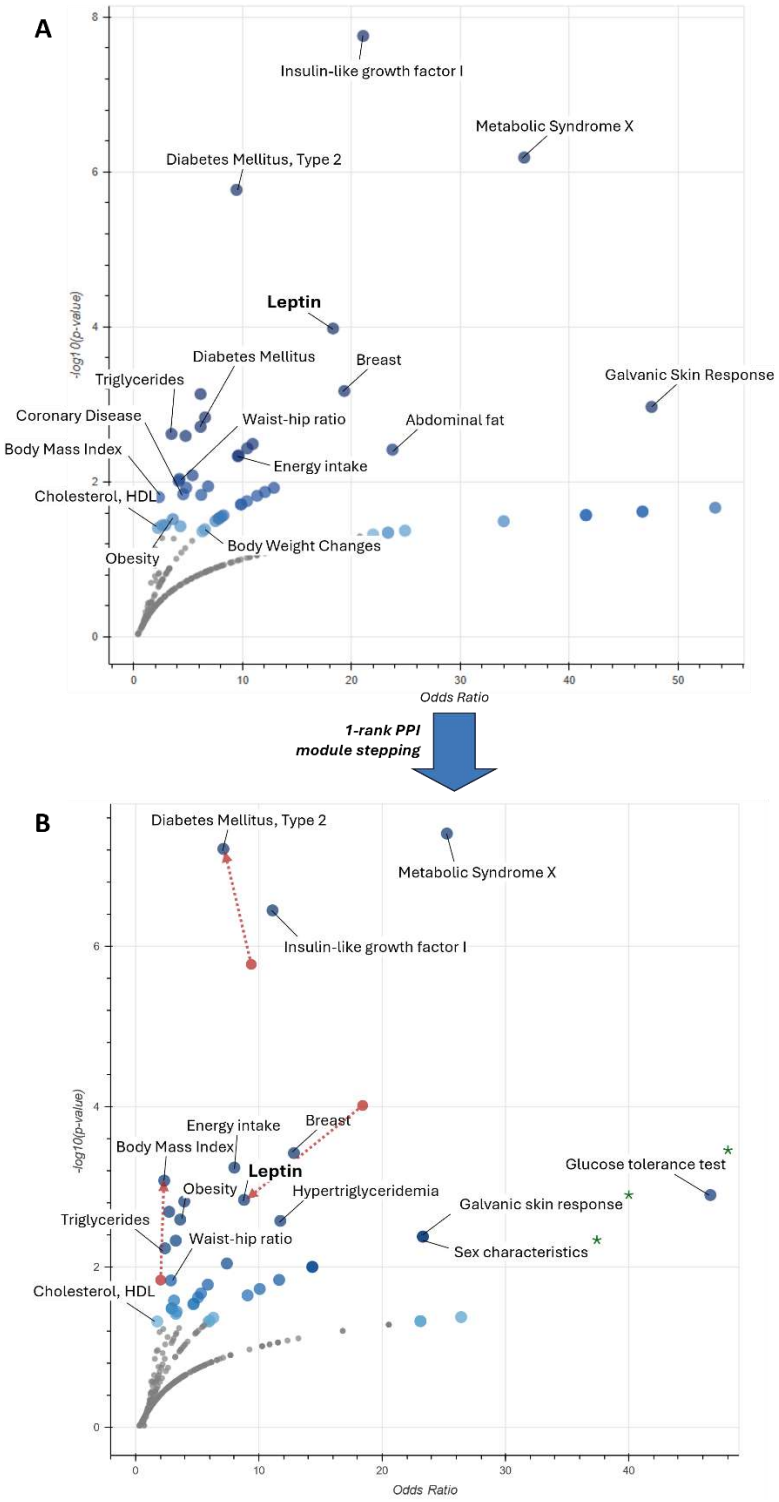


Figure 4. PPI network augmentation allows for biological term exploration. (A) PheGenI trait/disease enriched terms with the 57 leptin-related proteins. **(B)** PheGenI trait/disease enriched terms after 1-rank PPI stepping with 57 additional 1st-shell interactors, highlighting enriched and depleted terms after augmentation. Red arrows show the evolution of traits after the addition of interactors. Green asterisks (*) show new enriched traits.

On the other hand, the expanded PPI network was clustered into 9 PPI modules by k-means clustering, and each module was functionally annotated for biological pathways, ontological terms and diseases. Leptin-related proteins not interacting with the network were discarded. Each of the 9 resulting PPI modules governed specific biological functions, the central-one related to leptin and body-mass index, and others linked with either ciliary function and development, triglycerides and cardiovascular diseases, neuronal signaling and related disorders, or endocrine cell differentiation (**Figure 5**).

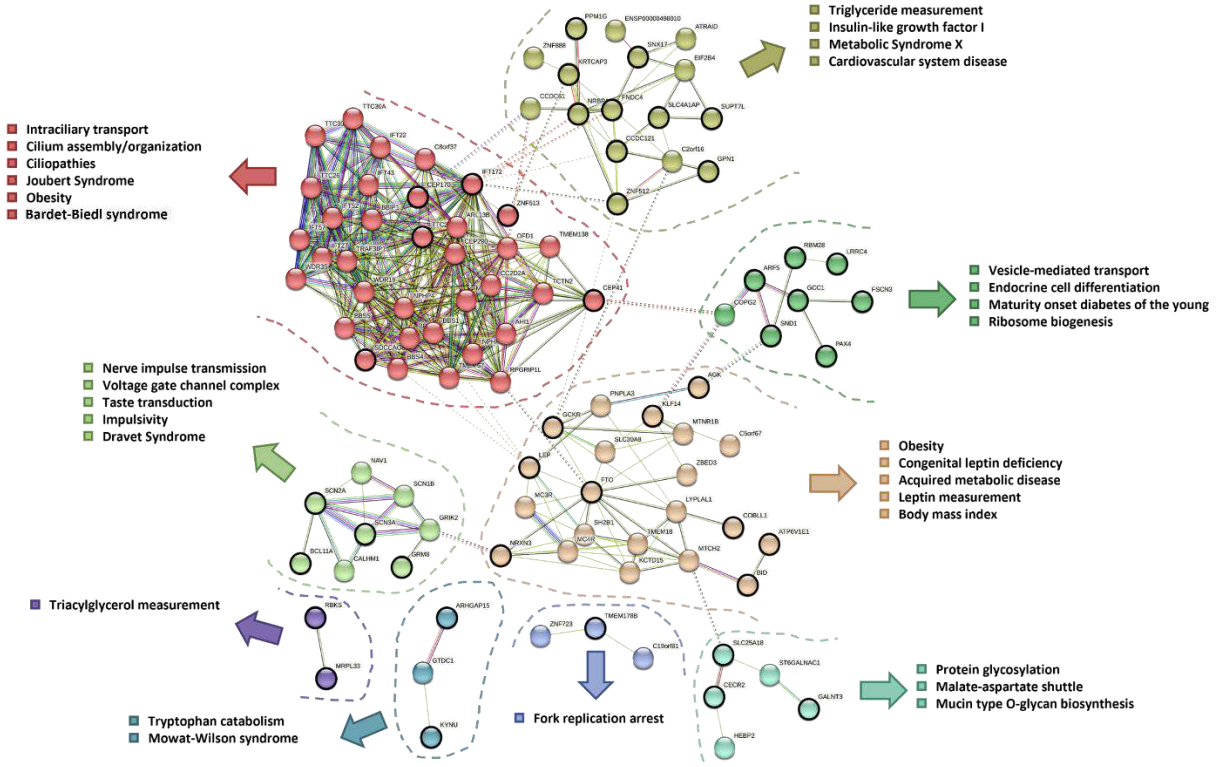


Figure 5. Augmented PPI network around leptin-related proteins. PPI network of the initial query proteins plus the 1st-shell interactors, clustered using the k-means algorithm. Leptin-related proteins are highlighted in black. Detached nodes (non-interacting proteins) were removed. Annotations for the main associated function, pathway, or disease accompany each cluster.

Discussion

Conducting the cross-traits analysis allowed for the identification of the phenotypes sharing unique genetic associations with serum leptin levels. Their uneven distribution among the 15 blocks of LD serves as a reminder that pleiotropy does not depend on block size, but rather on functional density. It also underscores the genomic architecture of blocks, where some regulatory elements hotspots (e.g., eQTLs, enhancers), while others are functionally dense regions – overlapping pleiotropic genes (e.g., GCKR, FTO). The disparity between blocks also highlights the Eurocentricity of GWAS arrays, which lead to under-tagging and limited coverage in African genomes (28).

These traits could be organized into 22 distinct categories impacting diverse functions and tissues, underlying leptin high pleiotropic effects in human. Specific categories showed depletion against background (brain-related, blood pressure, sleep and age-related), potentially indicating function specificity for distinct SNPs within general and/or subjective categories regrouping heterogeneous traits, or that leptin is functionally distant or indirectly linked with these traits and is hence less associated with them than random GWAS loci. On the other hand, high enrichment values were observed for diet, adipokines, addiction or obesity -related traits, suggesting shared underlying mechanisms with leptin levels and extended genetic overlap, pointing to potential functional regrouping and biological relevance of these traits in leptin's pleiotropy.

Functional consequences of these traits are highlighted by their effect directions with leptin serum levels, which largely confirmed physiological and mechanistical scientific literature in patients and animal models: homodirectional effects between leptin and body mass index (29) and obesity (30), and opposite effects with hyperlipidemia (29) and the long-guessed reverse balance with other adipokines such as adiponectin, which we confirm at a genetic level with this study.

We also provide a network of biological functions and pathways explaining these pleiotropic phenotypes. Upon expanding the PPI network around the leptin gene, we identified pathways linking leptin signaling and obesity to a cluster of proteins specialized in development and embryogenesis (especially ciliogenesis), a module of proteins governing metabolism regulation, a cluster implicated in endocrine regulation and cellular function, and a module of genes responsible for neuronal signaling and disorders. This echoes previous studies conducted on the implications of leptin in neuronal development, specifically that in the hypothalamus (31) where postnatal leptin acts as a neurotrophic factor and helps axonal growth within the arcuate nucleus, constructing the leptin-melanocortin neuronal pathway, which they will activate later in life to control satiety, energy intake and subsequent lipid metabolism.

A notable feature of network expansion is that it helps prioritize genes linked with the GWAS variants. This augmentation allows i) for a discard of the proteins that cannot interact with the PPI modules, which could stem from a wrong mapping of the corresponding genes in the GWAS studies, and ii) for a recovery – by adding functional interactors, till loss of "Leptin" phenotype – of genes that were not associated by GWAS or mapped in later stages. Losing the phenotype of interest (Leptin) marks the point beyond which augmentation becomes excessive, as new functional associations stop leading back to the main trait of the study.

The conducted study paves the way for further research aiming to elucidate obesity's physiopathology. A global correlation analysis by LD regression score between leptin and some of the identified cross-traits would hint at the overall genetic association between the traits. Moreover, causality between modifiable risk factors in a disease or phenotype and the individual's health outcomes can be deciphered with Mendelian randomizations (32). By conducting systematic MR between all possible cross-traits identified with leptin, one would be able to reconstruct the causal inference network, with the direction of causality between each linked trait. MR could be achieved using GWAS summary statistics for leptin and a number of traits and is also made possible with access to large cohorts like the UK BioBank, for which genomes and exomes are available for more than 500.000 individuals.

References

1. World Obesity Federation [Internet]. [cited 2025 May 20]. World Obesity Atlas 2023. Available from: https://www.worldobesity.org/resources/resource-library/world-obesity-atlas-2023

2. Guh DP, Zhang W, Bansback N, Amarsi Z, Birmingham CL, Anis AH. The incidence of co-morbidities related to obesity and overweight: a systematic review and meta-analysis. BMC Public Health. 2009 Mar 25;9:88.

3. Rohm TV, Meier DT, Olefsky JM, Donath MY. Inflammation in obesity, diabetes, and related disorders. Immunity. 2022 Jan 11;55(1):31–55.

4. Ni Mhurchu C, Rodgers A, Pan WH, Gu DF, Woodward M, Asia Pacific Cohort Studies Collaboration. Body mass index and cardiovascular disease in the Asia-Pacific Region: an overview of 33 cohorts involving 310 000 participants. Int J Epidemiol. 2004 Aug;33(4):751–8.

5. Bergström A, Pisani P, Tenet V, Wolk A, Adami HO. Overweight as an avoidable cause of cancer in Europe. Int J Cancer. 2001 Feb 1;91(3):421–30.

6. Pati S, Irfan W, Jameel A, Ahmed S, Shahid RK. Obesity and Cancer: A Current Overview of Epidemiology, Pathogenesis, Outcomes, and Management. Cancers (Basel). 2023 Jan 12;15(2):485.

7. Harvie M, Hooper L, Howell A h. Central obesity and breast cancer risk: a systematic review. Obesity Reviews. 2003;4(3):157–73.

8. Key TJ, Appleby PN, Reeves GK, Roddam A, Dorgan JF, Longcope C, et al. Body mass index, serum sex hormones, and breast cancer risk in postmenopausal women. J Natl Cancer Inst. 2003 Aug 20;95(16):1218–26.

9. Paracchini V, Pedotti P, Taioli E. Genetics of Leptin and Obesity: A HuGE Review. American Journal of Epidemiology. 2005 Jul 15;162(2):101–14.

10. De Matteis R, Puxeddu R, Riva A, Cinti S. Intralobular ducts of human major salivary glands contain leptin and its receptor. J Anat. 2002 Nov;201(5):363–70.

11. Bjørbaek C, Kahn BB. Leptin signaling in the central nervous system and the periphery. Recent Prog Horm Res. 2004;59:305–31.

12. Park HK, Ahima RS. Leptin signaling. F1000Prime Rep. 2014;6:73.

13. Tam V, Patel N, Turcotte M, Bossé Y, Paré G, Meyre D. Benefits and limitations of genome-wide association studies. Nat Rev Genet. 2019 Aug;20(8):467–84.

14. Comuzzie AG, Cole SA, Laston SL, Voruganti VS, Haack K, Gibbs RA, et al. Novel genetic loci identified for the pathophysiology of childhood obesity in the Hispanic population. PLoS One. 2012;7(12):e51954.

15. Meeks KAC, Bentley AR, Gouveia MH, Chen G, Zhou J, Lei L, et al. Genome-wide analyses of multiple obesity-related cytokines and hormones informs biology of cardiometabolic traits. Genome Med. 2021 Oct 7;13(1):156.

16. Wang X, Jia J, Huang T. Shared genetic architecture and casual relationship between leptin levels and type 2 diabetes: large-scale cross-trait meta-analysis and Mendelian randomization analysis. BMJ Open Diabetes Res Care. 2020 Apr;8(1):e001140.

17. Ortega-Azorín C, Coltell O, Asensio EM, Sorlí JV, González JI, Portolés O, et al. Candidate Gene and Genome-Wide Association Studies for Circulating Leptin Levels Reveal Population and Sex-Specific Associations in High Cardiovascular Risk Mediterranean Subjects. Nutrients. 2019 Nov 13;11(11):2751.

18. MacArthur J, Bowler E, Cerezo M, Gil L, Hall P, Hastings E, et al. The new NHGRI-EBI Catalog of published genome-wide association studies (GWAS Catalog). Nucleic Acids Res. 2017 Jan 4;45(D1):D896–901.

19. R: The R Project for Statistical Computing [Internet]. 2025 [cited 2025 Mar 14]. Available from: https://www.r-project.org/

20. Gel B, Serra E. karyoploteR: an R/Bioconductor package to plot customizable genomes displaying arbitrary data. Bioinformatics. 2017 Oct 1;33(19):3088–90.

21. Myers TA, Chanock SJ, Machiela MJ. LDlinkR: An R Package for Rapidly Calculating Linkage Disequilibrium Statistics in Diverse Populations. Front Genet [Internet]. 2020 Feb 28 [cited 2025 Mar 14];11. Available from: https://www.frontiersin.org/journals/genetics/articles/10.3389/fgene.2020.00157/full

22. Machiela MJ, Chanock SJ. LDlink: a web-based application for exploring population-specific haplotype structure and linking correlated alleles of possible functional variants. Bioinformatics. 2015 Nov 1;31(21):3555–7.

23. Szklarczyk D, Gable AL, Lyon D, Junge A, Wyder S, Huerta-Cepas J, et al. STRING v11: protein-protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets. Nucleic Acids Res. 2019 Jan 8;47(D1):D607–13.

24. Szklarczyk D, Kirsch R, Koutrouli M, Nastou K, Mehryary F, Hachilif R, et al. The STRING database in 2023: protein-protein association networks and functional enrichment analyses for any sequenced genome of interest. Nucleic Acids Res. 2023 Jan 6;51(D1):D638–46.

25. Chen EY, Tan CM, Kou Y, Duan Q, Wang Z, Meirelles GV, et al. Enrichr: interactive and collaborative HTML5 gene list enrichment analysis tool. BMC Bioinformatics. 2013 Apr 15;14:128.

26. Kuleshov MV, Jones MR, Rouillard AD, Fernandez NF, Duan Q, Wang Z, et al. Enrichr: a comprehensive gene set enrichment analysis web server 2016 update. Nucleic Acids Res. 2016 Jul 8;44(Web Server issue):W90–7.

27. Xie Z, Bailey A, Kuleshov MV, Clarke DJB, Evangelista JE, Jenkins SL, et al. Gene Set Knowledge Discovery with Enrichr. Current Protocols. 2021;1(3):e90.

28. Sawyer SL, Mukherjee N, Pakstis AJ, Feuk L, Kidd JR, Brookes AJ, et al. Linkage disequilibrium patterns vary substantially among populations. Eur J Hum Genet. 2005 May;13(5):677–86.

29. Kennedy A, Gettys TW, Watson P, Wallace P, Ganaway E, Pan Q, et al. The Metabolic Significance of Leptin in Humans: Gender-Based Differences in Relationship to Adiposity, Insulin Sensitivity, and Energy Expenditure*. The Journal of Clinical Endocrinology & Metabolism. 1997 Apr 1;82(4):1293–300.

30. Obradovic M, Sudar-Milovanovic E, Soskic S, Essack M, Arya S, Stewart AJ, et al. Leptin and Obesity: Role and Clinical Implication. Front Endocrinol (Lausanne). 2021;12:585887.

31. Bouret SG. Neurodevelopmental actions of leptin. Brain Res. 2010 Sep 2;1350:2–9.

32. Guo Z, Du H, Guo Y, Jin Q, Liu R, Yun Z, et al. Association between leptin and NAFLD: a two-sample Mendelian randomization study. Eur J Med Res. 2023 Jul 3;28:215.

Ten years of the Pasteur's Bioinformatics and Biostatistics Hub: achievements and perspectives

Hervé MENAGER[1], Damien MORNICO[1], Pascal CAMPAGNE[1], Elodie CHAPEAUBLANC[1], Claudia, CHICA[1], Julien GUGLIELMINI[1], Bertrand NÉRON[1], Natalia PIETROSEMOLI[1], Gaël MILLOT[1], Marie-Agnès DILLIES[1], Laurent ESSIOUX[1]

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

Corresponding Author: herve.menager@pasteur.fr

Keywords

Bioinformatics Platform, Research Support, Core facility, Computational Biology, Reproducibility

Background

Twenty-Twenty Five celebrates the 10th anniversary of the Hub of Bioinformatics and Biostatistics of the Institut Pasteur (henceforth "the Hub"). The rationale of the Hub's creation was to centralize bioinformaticians and biostatisticians who were previously scattered in different research lab Units within the campus, to mutualize support and create synergies. Since then, the Hub has become a central group on Pasteur campus, with the main objective of creating an environment of excellence in computational biology in support of research. It encompasses expert support on research projects, coaching scientists, training, methods and tools development and expert community building.

Results

Over the years, the Hub staff contributed to more than 800 projects (10% of which required more than 3 month-time effort), coming from 200 research units, covering all departments on campus, where multiple Hub's skills are often combined. It resulted in around 550 peer-reviewed publications, nearly 50 computational tools and 30 websites and portals deployed. The Hub created a dedicated PhD-training program and delivered trainings to more than 3000 participants. In addition, it has developed tailored courses and mentoring programs for Institutes that are part of the Pasteur Network. It has fully integrated into the bioinformatics French landscape, as exemplified by the co-organisation of the école de Bioinformatique IFB (EBAII) and the workshop on single-cell data analyses (SincellTE) since 2018, as well as by the establishment of an open-access Galaxy instance designed to support the

community in their bioinformatics analyses and the numerous contributions to the Tools platform of the ELIXIR european infrastructure.

The Hub adopted a hub-and-spoke structure with currently more than 40 core Hub members being distributed in five poles and 30 engineers being detached in research/technology units and in the transmissible disease surveillance labs ("Centre Nationaux de Reference (CNR)") under Pasteur's responsibility. Such detachments are renewable five-year assignments and research engineers (the spokes) dedicate 20% of their time to the Hub. Sets of tools, meetings and governance rules have been set up to address at best the diversity of scientists needs: weekly open-desks, a web-based in-house project management tool to handle short requests, coaching and small projects, and a steering committee to assess and grant "long projects" (>3 month-time effort) and detachment renewals. To ensure methodological developments, expertise growth and collaborations with the bioinformatics communities and networks, 20% of each engineer's time is dedicated to innovation work. In addition, The Hub is affiliated to the computational biology department to create a stimulating scientific environment and foster collaborations in methodology development and assessment.

Conclusion

The initial promises and expectations raised at the Hub's creation have been fulfilled and the Hub does impact Pasteur's research. Importantly, it must be underlined that such a large and diverse group offers the opportunity for computational biologists to quickly address questions outside their area of expertise, learn from experts and grow their expertise. It prevents expertise silos and the "miss/mr does it all" situation of isolated bioinformaticians. Finally, detachment periods with affiliation to the Hub offer more flexible career experiences to the computational biology research engineers' community.

Beyond the rich human endeavour the ten years represent, we will share our experience in setting up and leading a large computational biology platform. We will discuss methods to overcome the resistance to a complex support model, and key lessons regarding governance, operational model, methodology development and outreach in such a structure. We will highlight challenges such as career development, thriving through diversity, time-management, sustained innovation in an ever-faster evolving field and finally our perspectives for the next ten years!

Explainable AI for Marine Ecological Quality Prediction: Integrating Microbiome Data, Metadata, and Diversity

Houria BRAIKIA[1], Sana BEN HAMIDA[1,2] and Marta RUKOZ[1,2]

[1] LAMSADE, Paris Dauphine University- PSL, Pl. du Maréchal de Lattre de Tassigny, 75016 Paris, France
[2] SEGMI, Paris Nanterre University, France

Corresponding author: houria.braikia@dauphine.psl.eu

Keywords Marine Biodiversity Monitoring, Ecological Quality, SHAP, Alpha Diversity

Abstract *Assessing ecological quality (EQ) is crucial for marine biodiversity monitoring. With the advent of High-Throughput Sequencing technologies, metabarcoding has enabled large-scale microbial community analysis through Operational Taxonomic Unit (OTU) tables, providing an alternative for EQ assessment. Machine learning (ML) models have been successfully applied for this task, but they often treat microbial abundance as the sole predictor, overlooking environmental meta-data (e.g., pH, salinity, temperature) and diversity indices (alpha and beta diversity). This study integrates metadata and diversity indices into an explainable ML framework for EQ prediction. Using SHapley Additive Explanations (SHAP), we assess the contribution of these features to model predictions across five genetic markers (V1V2, V3V4, V4, 37F, and V9). Our results highlight marker-dependent feature importance, demonstrating that while OTU-based models remain dominant, incorporating metadata improves accuracy for certain markers. This work enhances interpretability in AI-driven biomonitoring, fostering more reliable marine ecosystem assessments.*

Introduction

The integration of machine learning (ML) for predicting the Biotic Index (BI) to assess the Ecological Quality (EQ) of marine environments using environmental DNA (eDNA) metabarcoding data represents a significant advancement in biomonitoring. This assessment often involves computing BI values, which can be translated into five ecological quality classes ranging from "very good" to "very bad". BI values can be calculated using benthic macroinvertebrate data or through metabarcoding data obtained from high-throughput amplicon sequencing of eDNA. The former method is time-consuming and requires extensive taxonomic expertise, while the latter relies on reference databases, limiting its applicability to sequences that have known taxonomic or ecological annotations.

Recent studies have proposed an alternative approach that uses supervised machine learning (SML) to generate predictive models for BI values or EQ classes directly from eDNA data. The first study by [1] in this category tested two SML approaches to predict BI values by focusing on specific taxonomic groups, using benthic foraminifera as features to infer four commonly used BI values for benthic monitoring (AZTI Marine Biotic Index (AMBI) [2], Indicator Species Index (ISI) [3], Norwegian Sensitivity Index (NSI) [3], and Norwegian Quality Index 1 (NQI1) [4]). Their results demonstrated that SML approaches could provide accurate BI predictions, reducing or even eliminating the time and cost constraints associated with morphology-based assessments. In a follow-up study ([5]), the same authors trained Random Forest (RF) models using five different genetic markers—eukaryotic markers (V1V2, V4, and V9), ribosomal bacterial markers (V3V4), and foraminifera markers (37F) to evaluate

how predictive accuracy varies across markers. Their findings indicated that all tested markers produced reliable predictive models, outperforming conventional taxonomic classification approaches. Additionally, the study suggested that the predictive performance of the 37F marker was slightly lower than that of the eukaryotic and bacterial markers, confirming that biomonitoring models perform better when a broader taxonomic spectrum is used as features.

More recently, [6], introduced a simplified approach to predicting EQ, demonstrating that directly predicting EQ classes yielded better results than focusing on individual BI values. This finding reinforces the growing recognition of ML's potential in biomonitoring, particularly for EQ assessment.

However, none of the previously mentioned studies have examined how environmental conditions contribute to ecological quality. Building on [6]'s work, this study integrates model explainability to explore the influence of environmental variables, the role of diversity, and how ML algorithms capture these relationships. Feature importance methods quantify each variable's contribution to predictions, with global methods ranking features overall and local methods assessing their impact on specific cases.

We implement an explainable artificial intelligence (XAI) framework using microbiome data, metadata, and alpha diversity indices to enhance marine EQ predictions. Specifically, we apply SHapley Additive Explanations (SHAP) to identify key variables driving model decisions. By improving interpretability, this study strengthens the reliability and transparency of ML-based ecological assessments, facilitating better-informed marine monitoring efforts.

Background

Alpha Diversity Indices in Marine Microbiome Analysis

Biodiversity assessment depends on diversity metrics to quantify species richness and distribution. In marine ecology, these metrics, or indices, characterize microbial community composition under varying environmental conditions, offering insights into ecosystem health. Diversity indices are mathematical measures that summarize species-abundance distribution within a community as a single value, offering a snapshot of diversity and its fluctuations over time ([7]). The three primary diversity categories are Alpha Diversity (measure within-sample), Beta Diversity (measure between-sample), and Gamma Diversity (measure in landscape-level). Alpha diversity, one of the most widely used metrics for characterizing communities at a local scale, consists of two main components: species richness, which represents the number of different species, and equitability, which measures the uniformity of species abundances ([8]).

Species Richness (S) quantifies species count but does not consider distribution. To address this, evenness indices are used.

We begin by introducing some notations that are used in the formulas of the indices introduced later on. Consider S as the species richness, and let $p_i = \frac{n_i}{N}$ the proportional abundance (or percentage abundance) of the i-th species present, where n_i is the number of individuals for species i, and N is the total number of individuals counted, across all species.

Shannon-Wiener Index (H) ([9]) quantifies alpha diversity by considering both species richness and evenness. A higher Shannon index suggests evenly distributed species, while lower values indicate the dominance of a few species ([8]). It is computed as follows 1:

$$H = -\sum_{i=1}^{S}(p_i \cdot \log_2(p_i)) \tag{1}$$

Pielou's Evenness Index (P) ([10]) measures species distribution independence from richness, making it useful for dominance comparisons ([8]). It is given by 2:

$$P = \frac{H}{\log_2(S)} \tag{2}$$

Values range from 0 (one species dominates) to 1 (equal distribution).

Simpson's Index (D) ([11]) estimates the probability that two randomly chosen individuals belong to the same species, giving more weight to abundant species ([8]). It is defined as 3:

$$D = 1 - \sum_{i=1}^{S}(p_i^2) \tag{3}$$

Unlike the Shannon index, Simpson's Index is less sensitive to rare species, making it more suited for habitat comparisons.

Multivariates Analysis of Microbiome Data

Multivariate analysis refers to a collection of statistical techniques that examine three or more variables simultaneously, in order to identify or clarify the relationships between them. Unlike univariate analysis, which focuses on a single observation or variable, multivariate analysis acknowledges the complexity of real-world phenomena, where multiple factors influence outcomes ([12]).

We may investigate individual measurements, in simple cases, using measures of location and dispersion, or explore relationships between two variables using bivariate analysis. However, most datasets involve numerous variables, and understanding the interrelationships between them requires a multivariate approach. This type of analysis is applicable to both metrical and categorical data, and offers a variety of methods to uncover meaningful patterns and associations ([13]).

Using correlation and Canonical Correspondence Analysis, we will explore the relationships between environmental variables (metadata) and microbiome data, while SHAP will be used to examine the connection between ecological quality, environmental factors, microbiome, and diversity.

Correlation and Canonical Correspondence Analysis Correlation measures the strength and direction of the linear relationship between two variables, quantifying the extent to which they change together. While the linear relationship between each pair of independent variables in the metadata can be measured by correlation, a more insightful approach to explore the multivariate relationships between community composition and multiple explanatory environmental variables is Canonical Correspondence Analysis (CCA).

CCA is a multivariate analysis technique designed to relate community composition directly to environmental variation. It works by imposing the restriction that ordination axes be linear combinations of environmental variables, allowing community variation to be directly linked to environmental factors. The environmental variables can be either quantitative or nominal, and as many axes as there are environmental variables can be extracted ([14]).

CCA is particularly useful for analyzing data when species exhibit bell-shaped response curves to environmental gradients, making it more appropriate for examining the relationship between community composition and environmental variables than other methods like canonical correlation analysis. The technique produces an ordination diagram in which species and sites are represented as points, and environmental variables are shown as vectors. This diagram helps visualize the patterns of variation in community composition explained by environmental variables and identifies the "centers" of species distributions along each environmental gradient. CCA has been effectively applied to various ecological studies, such as hunting spiders, dyke vegetation, and algae along pollution gradients ([14]).

Explainability in Machine Learning and SHAP Understanding why a model makes a specific prediction is often just as important as how accurate that prediction is—especially in fields where interpretability matters. Complex models like ensemble methods or deep learning can perform very well with large datasets, but they're often difficult to understand. This creates a tension between achieving the highest performance and maintaining a model that users can understand. In response to this issue, various methods have been proposed to help interpret the predictions of complex models. However, it is often unclear how these methods relate to one another and in which contexts one method may be more suitable than another ([15]).

To address this challenge, SHapley Additive Explanations (SHAP) ([15]) was introduced as a unified framework for interpreting model predictions. SHAP assigns an importance value to each feature based on its contribution to the model's output. This method is grounded in cooperative game theory and introduces a novel class of additive feature importance measures, which is theoretically proven to have desirable properties. These properties unify several existing methods and ensure that SHAP provides both global and local interpretations of model behavior. SHAP's distinctive advantage lies in its ability to provide consistent, interpretable feature importance, while also offering computational efficiency compared to earlier approaches ([15]).

In this study, we apply SHAP to explain the predictions of the Random Forest (RF) ([16]) model, which is an ensemble method that constructs multiple decision trees through bootstrapping and random feature selection. RF predictions are made by averaging (for regression) or majority voting (for classification) across the trees. SHAP is used to assess the influence of operational taxonomic units (OTUs), metadata, and diversity indices on EQ predictions. This helps make our machine learning models more transparent, offering clearer insights into how different factors—like environmental conditions and microbial diversity—affect ecological assessments.

Material and Data Analysis

Material

In this study, we analyze five datasets that are publicly available on zenodo [4] from [5] study, corresponding to five genetic markers used in marine biodiversity research: Eukaryotic markers (V1V2, V4, and V9), Ribosomal bacterial marker (V3V4) and Foraminifera marker (37F). The dataset is divided into two types. The first type includes metadata files that list the sample names along with four BIs, namely AMBI, ISI, NSI, and NQI1. Other columns in these metadata files contain environmental data, such as sampling location, station, grab, distance from the farming cage, depth, and pH. The second type consists of five files, each representing an OTU-sample matrix for a different marker. These files contain the sample names and the corresponding abundance of each OTU in those samples. This data serves as the basis for analyzing the relationships between OTU abundance and environmental factors.

The pipeline code of this study is available on zenodo [5].

Alpha diversity Analysis

To better understand the microbial community structure across the five datasets (V1V2, V3V4, V4, 37F, V9), we conducted alpha diversity analyzes. The results for each dataset are visualized in Figure 1, which provides a comparative box plot highlighting the variation of the alpha diversity indices.

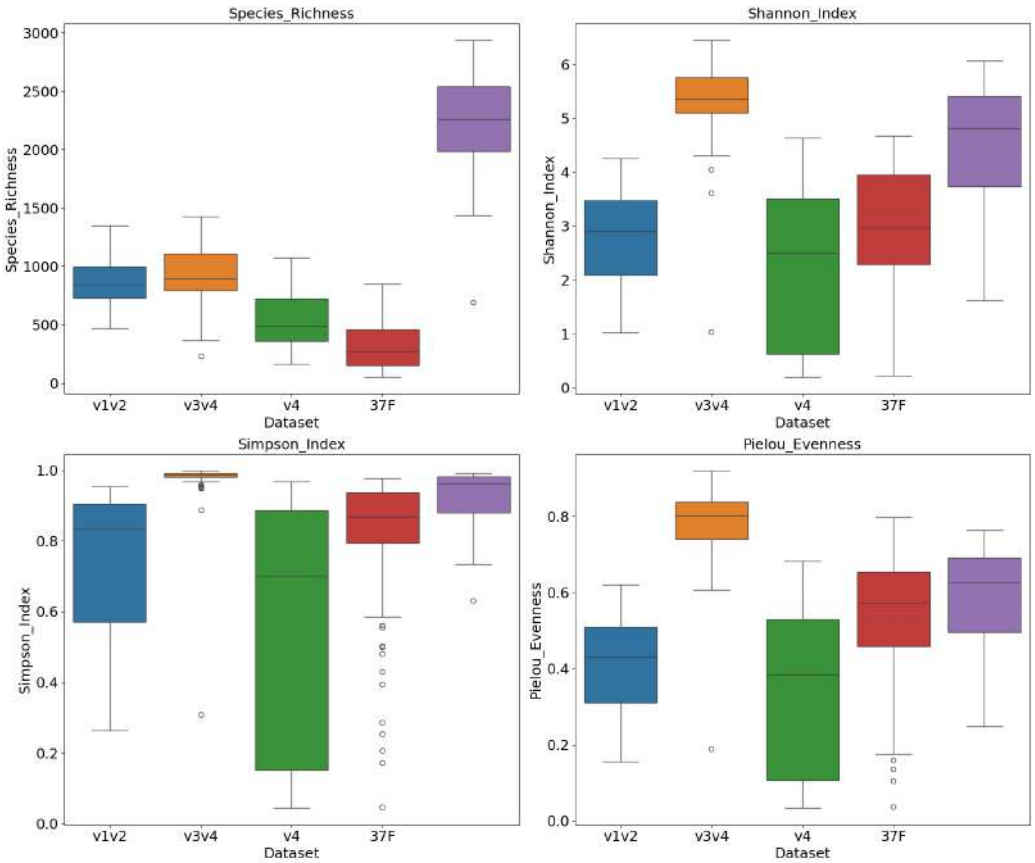


Fig. 1. *Box Plot Comparison of Alpha Diversity Indices Across Datasets.*

4. DOI:10.5281/zenodo.1286476
5. DOI:10.5281/zenodo.15238612

— **Species Richness :** The species richness box plots reveal significant differences in the number of species detected by each marker. The medians of V1V2 and V3V4 are close, both below 1000, reflecting a moderate number of species captured in these datasets. In contrast, the V4 dataset has an even lower median, below 500, indicating a reduced species richness, while 37F shows the lowest median overall, highlighting its limited capacity to capture species. On the other hand, V9 stands out with the highest median, ranging between 2000 and 2500, suggesting that this marker captured the most species. Outliers are visible in the V3V4 and V9 datasets, indicating some extreme values outside the typical range.

  When examining the interquartile range (IQR), V9 shows the largest IQR, reflecting the greatest variability in species richness across its samples. The other datasets display more similar IQRs, indicating a more consistent distribution of richness values. The whiskers further clarify the spread: V1V2 has balanced, moderately long whiskers, while V3V4 and V9 show longer lower whiskers, suggesting a wider range of lower values. In contrast, V4 and 37F have longer upper whiskers, indicating greater variation in higher richness values. These trends are consistent across the box plots and highlight key differences in species richness captured by each marker.

— **Shannon Index :** The Shannon index, which accounts for both species richness and evenness, reveals a relatively consistent pattern across the datasets. Among them, V3V4 stands out with the highest median value, suggesting it captures the greatest overall species diversity. This dataset also includes a notable outlier, indicating the presence of an exceptionally low diversity value.

  V1V2 displays whiskers of roughly equal length, with a median close to those of V4 and 37F, implying a similar central tendency in diversity across these three markers. The V4 dataset, with a median just below 3, has the widest IQR, reflecting greater variability in diversity among its samples. The lower half of the box is longer, indicating more spread in samples with lower diversity.

  Likewise, 37F exhibits a longer lower whisker than the upper one, highlighting greater variability at the lower end of its diversity distribution. As for V9, it has a higher median than V1V2, V4, and 37F, and shows a longer lower whisker as well—again pointing to more dispersion in samples with lower diversity scores.

— **Simpson Index :** The Simpson index, which emphasizes dominance and evenness in species distribution, supports earlier observations. V9 and V3V4 show values very close to 1, suggesting a highly even distribution with no dominant species. V3V4 has the highest median and an extremely small IQR, indicating both high evenness and low variability across samples. In contrast, V4 displays the lowest median and the largest box, with a longer lower half, pointing to lower evenness and greater variation among samples—suggesting that some species may dominate.

  V1V2 and 37F have close medians, indicating comparable diversity levels. The IQR is particularly narrow for V3V4, showing consistent values across its samples. Most datasets contain outliers. The longer lower whiskers in V1V2, 37F, and V9 suggest greater dispersion among the samples with lower diversity scores.

— **Pielou's Index :** Finally, Pielou's evenness index shows that V3V4 has the highest evenness values and median, reflecting a uniform distribution of species across its samples. In contrast, V1V2 and V4 exhibit the lowest evenness and the greatest variability, suggesting a more uneven distribution where certain species may dominate.

As with the other indices, V4 displays a large box, with the lower portion being larger, indicating more variation in samples with low evenness. 37F and V9 have similar medians and show longer lower whiskers, reflecting greater variability among samples with lower evenness values.

Multivariate Analysis

Exploring Relationships Among Environmental Condition Variables We start to explore the relationship between environmental variables and BIs. To achieve this, we use the metadata table, which includes columns such as Sample Names, Locality, Station, Grab, Distance from the Cage, Depth, pH, AMBI, NSI, ISI, and NQI1. Depth, pH, AMBI, NSI, ISI, and NQI1. To ensure clarity and focus in our analysis, we retained only the columns relevant to our research objectives: Sample Names, Distance from the Cage, Depth, pH, AMBI, NSI, ISI, and NQI1. We also addressed any null values to maintain the integrity of the dataset. We then calculated the correlation between each environmental variable and each BI.

We calculated the correlation between each environmental variable and each BI. The heatmap in Figure 2 shows a strong correlation between Distance Cage and the four BIs, with coefficients exceeding 0.7 (a negative correlation with AMBI and positive correlations with the others). Additionally, there is a relatively strong correlation between Depth and the four BIs, with coefficients above 0.5, while the correlation between pH and the four BIs is weaker, with coefficients above 0.4. The corresponding p-values showen in table 1 for all Pearson tests are very small, indicating that these correlations are statistically significant.

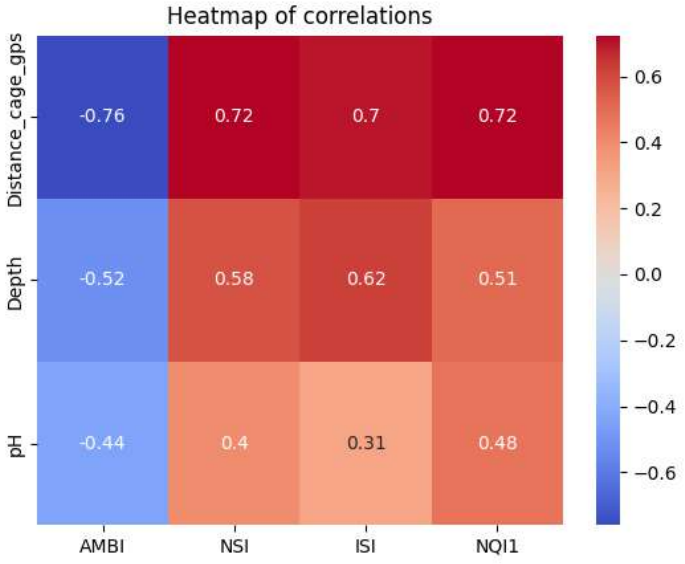


Fig. 2. Heatmap showing correlations between environmental variables and BIs.

Variable	AMBI	NSI	ISI	NQI1
Distance from Cage	4.176×10^{-26}	9.462×10^{-23}	1.912×10^{-20}	1.850×10^{-22}
Depth	2.037×10^{-10}	3.929×10^{-13}	1.258×10^{-15}	2.934×10^{-10}
pH	1.594×10^{-7}	2.253×10^{-6}	3.012×10^{-4}	7.948×10^{-9}

Tab. 1. P-values of Pearson correlation tests between environmental variables and ecological quality indices.

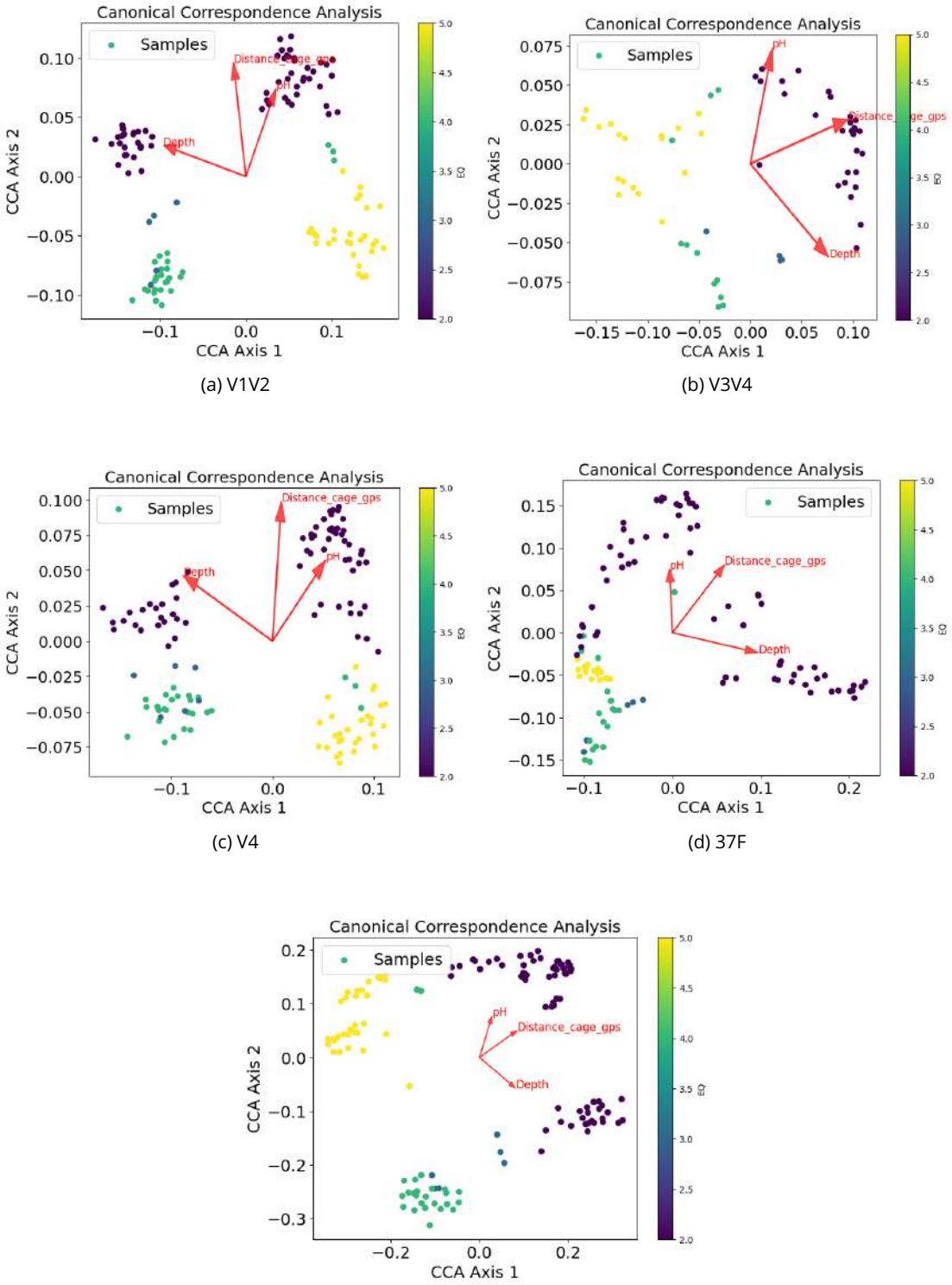
Canonical Correspondence Analysis of Environmental Variables and Ecological Quality To further explore the relationship between environmental variables and EQ, we applied CCA. This analysis was performed using three environmental variables including pH, distance from the cage, and depth. The EQ variable used here was inferred from AMBI scores based on the threshold values defined in Borja et al. [2]. We then visualized the results by projecting the samples onto a 2D space defined by the first two axes of the CCA (Figure 3). To examine how EQ is distributed, the samples were color-coded according to their EQ classes derived from AMBI values, allowing us to assess the influence of environmental factors on ecological status.

The resulting plot revealed that the samples grouped primarily by their EQ classification, with clear clusters. The arrows representing the environmental variables showed the direction of influence for each factor. Notably, pH, distance from the cage, and depth all demonstrated a clear relationship with the samples' ecological quality. The arrows for pH and depth point toward areas with higher ecological quality, indicating that higher pH levels and deeper sampling depths are associated with better EQ status. In contrast, the distance-from-cage variable shows a negative relationship: samples located closer to the cage tend to have lower ecological quality. These patterns suggest that environmental conditions like higher pH and greater depth may support healthier ecosystems, while proximity to aquaculture structures could be linked to ecological degradation.

These results from the CCA reinforce the conclusions drawn from the correlation analysis, highlighting the importance of environmental variables in influencing marine ecosystem health and providing a visual representation of how these factors contribute to variations in ecological quality across samples.

Results and Explainability In order to explore the relationships between explanatory variables—diversity, environmental variables, microbiome data—and the explained variable, ecological quality, we implemented an SML pipeline that takes the explanatory variables as input to predict the explained variable.

This pipeline is organized into three steps. First, the data in OTU tables, along with their metadata, are preprocessed to generate the training/testing datasets. The OTU tables were normalized using the Trimmed Mean of M-values (TMM) method, then reduced using Singular Value Decomposition (SVD) to five dimensions to adress the issue of the curse of dimensionality (a high number of features relative to the number of samples) and avoid overfitting. As in [6], SVD and TMM proved to be the best approach for these data. The number of components to retain for SVD is determined by maximizing both the variance between features and the accuracy of the RF model while using the minimum number of components possible.



(a) V1V2

(b) V3V4

(c) V4

(d) 37F

(e) V9

Fig. 3. Canonical Correspondence Analysis plot ordination plots for five datasets.

For the environmental variables, we retained "Distance from cage", "Depth", and "pH", which were scaled using "StandardScaler()" python function. As for the explained variable, following [6], we kept the AMBI index, which was converted into EQ classes.

The second phase is the learning step. We trained our RF model, with 200 trees, on three tasks: (1) OTU data extended with metadata and diversity indices, (2) metadata only, and (3) diversity indices only. These RF models are used to generate predicted EQ classes for each sample in the test set.

The third step quantifies the contribution of all explanatory variables in EQ predictions using SHAP values and RF feature importance. The objective is to determine whether richness, evenness, metadata, or community composition shifts play a more significant role in ecosystem health.

The models are evaluated using the kappa statistic, and the results are presented in Table 2 for each dataset. This table first presents results from the literature, specifically those of [5] and [6], respectively, followed by our results for extended data, metadata, and diversity data. The best results are highlighted in bold.

Cordier et al. first predicted AMBI values, then converted them into EQ classes before calculating the Kappa score to assess the agreement between predicted and reference EQ classes. Unlike our approach, they did not apply dimensionality reduction. Braikia et al., on the other hand, directly predicted EQ classes using only normalized and reduced OTU tables (non extended data).

Our results indicate that the effect of adding metadata and diversity indices varies across markers. For V3V4 and V4 markers, our approach outperforms Braikia et al.'s results. When compared to Cordier et al., our method surpasses their results for the V1V2, V4 and 37F markers. These findings suggest that the contribution of metadata and diversity indices depends on the specific marker, affecting the predictive power of the models in different ways.

Upon further analysis, we observe that predictions based solely on metadata perform relatively well. This suggests that the environmental or contextual information captured in the metadata is valuable for predicting ecological quality. On the other hand, predictions using only diversity indices show bad performance, indicating that diversity alone may not provide sufficient information to accurately assess ecological quality.

Tab. 2. Kappa results comparison

Markers	Literature Kappa	Non extended data	Extended data	Meta data	Diveristy data
V1V2	0.866	**0.956**	0.889	0.889	0.389
V3V4	**0.918**	0.834	0.863	0.713	0.514
V4	0.877	0.916	**0.928**	0.893	0.552
37F	0.832	**0.889**	0.859	0.816	0.771
V9	**0.927**	0.881	0.837	0.836	0.468

To verify these hypotheses and better understand the role of metadata and diversity indices in the prediction of EQ, we analyzed feature importance using both RF and SHAP values. The feature importance histograms, first generated by the RF model and then by the SHAP method for each dataset, are displayed in Figure 4.

The order of feature importance is generally consistent between RF and SHAP, except for some of the least important attributes. Otherwise, the ranking of important features remains similar across all markers. We note that the most influential features For markers V1V2, V3V4, V4, and V9 include OTU

attributes (reduced) and the meta-variable "Distance to the cage.", and that the diversity indices contribute minimally to the prediction of EQ. In particular, for V1V2, diversity indices do not significantly explain ecological quality, a result confirmed by the kappa value when testing predictions using only diversity data (table 2).

However, the case of 37F stands out. According to both SHAP and the feature importance scores from RF, diversity indices play a much more significant role in explaining EQ compared to OTU data. This finding suggests that, for 37F, diversity indices may be more informative than OTU-based features in predicting EQ.

As demonstrated in our results, the performance of V4 was enhanced by integrating environmental variables and diversity measures, followed by V1V2, which showed the second-best results. To further explore the impact of each variable and identify key drivers of good and poor ecological quality, we analyzed SHAP values for the "Very Good" and "Very Bad" classes in both datasets (Figure 5).

Focusing on the five most influential variables, we observe that in V1V2, high values of environmental variables strongly drive the classifier toward good quality. Similarly, the reduced OTU components 2 and 1 exhibit a pattern where low values contribute to good quality, whereas OTU component 4 has the opposite effect (Figure 5 (a)). Conversely, for the Very Bad class (Figure 5 (b)), SHAP values confirm these trends: high values of depth and distance from the cage push the classifier towards bad quality, while pH appears to have a mitigating influence.

For V4, a similar trend is observed. The five most important variables (Figure 5 (c)(d)) include environmental factors and reduced OTU components. As in V1V2, higher values of distance from the cage are associated with good ecological quality, whereas lower values tend to indicate poorer quality.

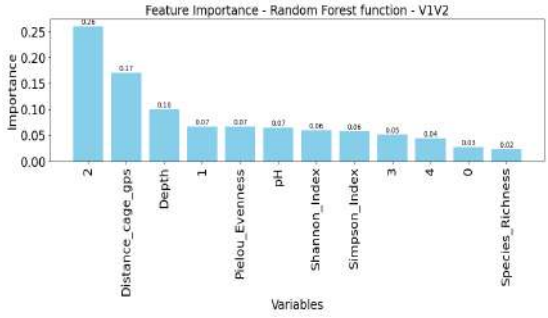
These findings highlight the strong influence of environmental parameters and microbial diversity on ecological quality classification, reinforcing the relevance of integrating these variables into predictive models.

Conclusion

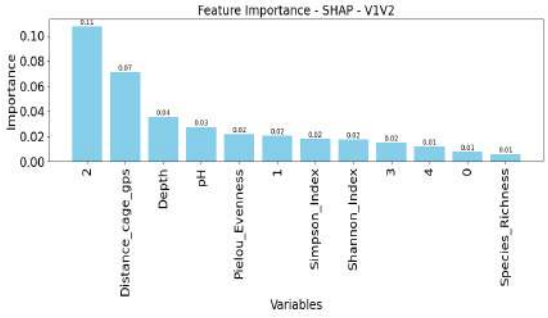
The goal of this study was to examine the contribution of environmental, diversity, and microbiome variables in predicting EQ. We assessed this by evaluating an RF model on these data and, to verify and understand the model's predictions, we used SHAP.

We developed an explainable ML framework for predicting marine EQ from eDNA metabarcoding data. Our pipeline, based on SVD and RF, was tested across five genetic markers (V1V2, V3V4, V4, 37F, and V9), and we evaluated the model on three tasks: (1) OTU data extended with metadata and diversity indices, (2) metadata only, and (3) diversity indices only. For the environmental variables, we retained "Distance from cage," "Depth," and "pH," while for diversity indices, we calculated richness, Shannon, Simpson, and Pielou. We then applied SHAP to measure the contribution of different features to the model's predictions.

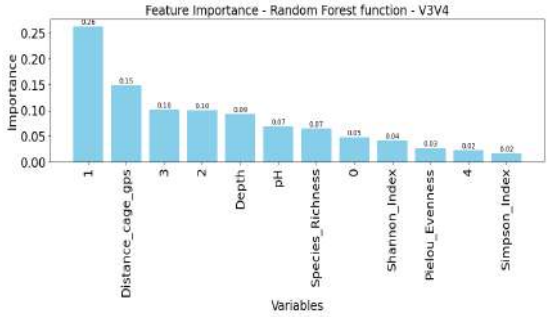
The results showed that predictive performance varied by genetic marker, with OTU-based features consistently being the most important. The contribution of metadata was significant, while diversity indices showed mediocre performance, differing substantially across datasets. These additional



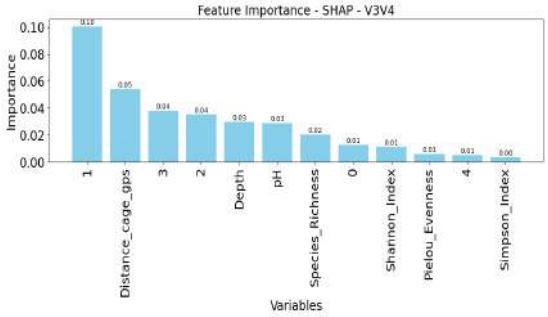
RF Feature importance for V1V2



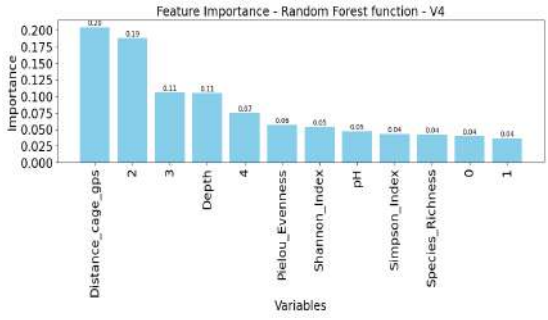
SHAP Feature Importance for V1V2



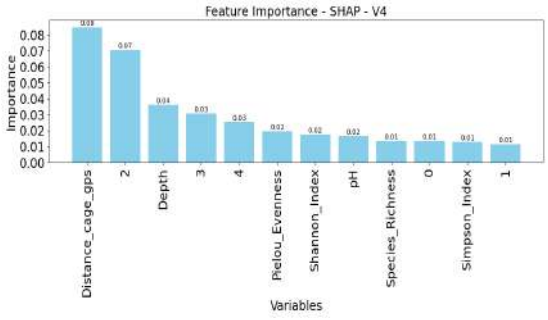
RF Feature importance for V3V4



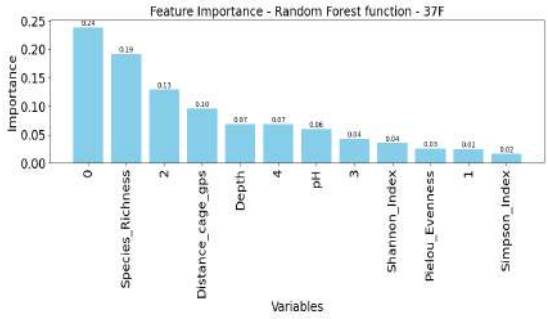
SHAP Feature Importance for V3V4



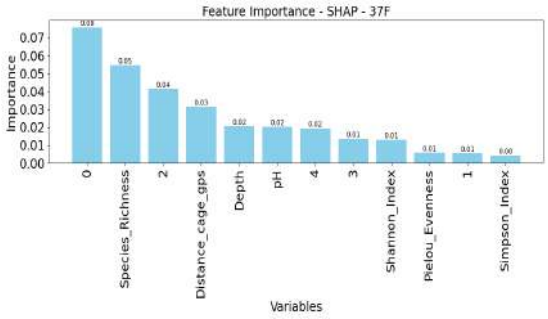
RF Feature importance for V4



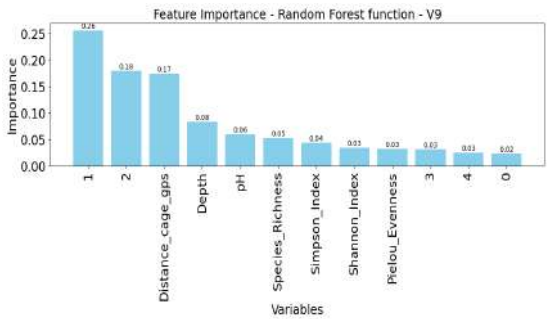
SHAP Feature Importance for V4



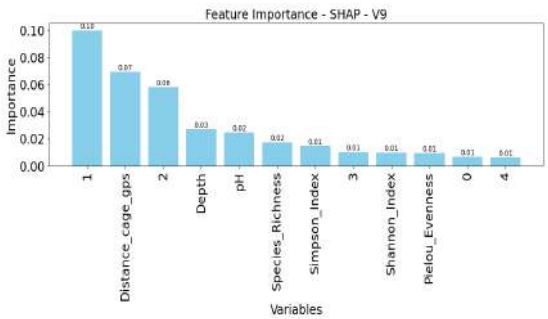
RF Feature importance for 37F



SHAP Feature Importance for 37F

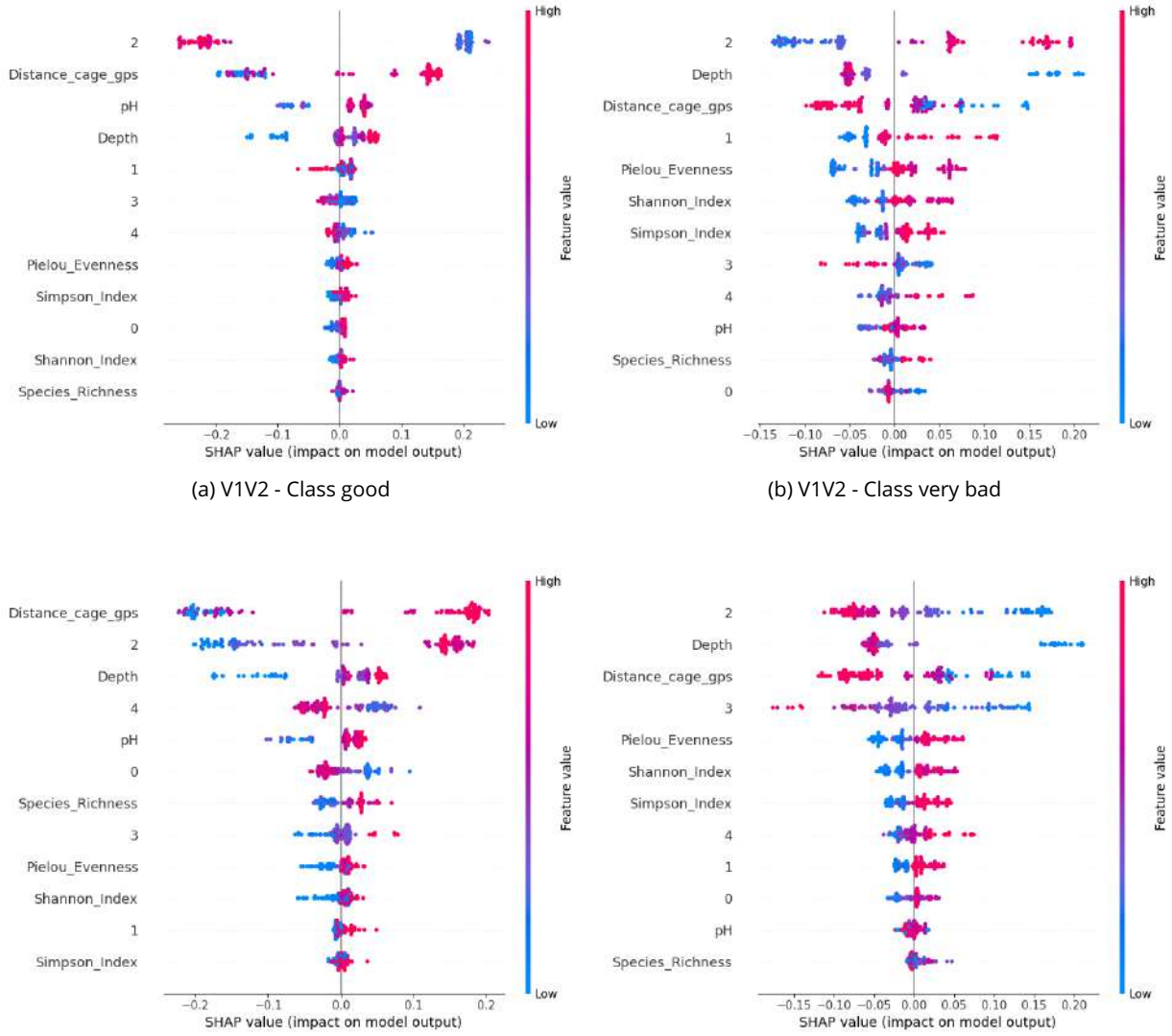


RF Feature importance for V9



SHAP Feature Importance for V9

Fig. 4. *Random Forest and SHAP Feature Importance in EQ Prediction Across Five Datasets (V1V2, V3V4, V4, 37F, and V9). The numbers in the x-axis represent the reduced OTU data.*



(a) V1V2 - Class good

(b) V1V2 - Class very bad

(c) V4 - Class good

(d) V4 - Class very bad

Fig. 5. Feature Impact on predicting "Good" and "Very Bad" classes (Datasets V1V2 & V4) according to SHAP: Environmental variables (Depth, pH, Distance from Cage) and OTU reduced to 5D.

features improved EQ predictions for V1V2, V3V4 and V4, had minimal impact on V9, and negatively affected the predictions for 37F.

SHAP and RF feature importance analyses confirmed these trends, revealing that OTUs and metadata—particularly distance from the cage—played a dominant role in EQ predictions, except for 37F, where diversity indices were more informative.

These findings emphasize that the predictive power of metadata and diversity indices is marker-dependent, highlighting the need for marker-specific feature selection strategies in future studies.

References

[1] Cordier T, Esling P, Lejzerowicz F, Visco J, Ouadahi A, Martins C, et al. Predicting the ecological quality status of marine environments from eDNA metabarcoding data using supervised machine learning. Environmental Science and Technology. 2017 Jun;51.

[2] Borja A, Franco J, Pérez V. A marine biotic index to establish the ecological quality of soft-bottom benthos within European estuarine and coastal environments. Marine Pollution Bul-

letin. 2000;40(12):1100-14.

[3] Rygg B, Norling K. REPORT SNO 6475-2013 Norwegian Sensitivity Index (NSI) for marine macroinvertebrates, and an update of Indicator Species Index (ISI). 2013 Jan.

[4] Rygg B. Developing indices for quality-status classification of marine soft-bottom fauna in Norway. Norsk institutt for vannforskning; 2006. Accepted: 2014-08-01T10:50:25Z ISSN: 1894-7948 Publication Title: 33 https://niva.brage.unit.no/niva-xmlui/handle/11250/213219.

[5] Cordier T, Forster D, Dufresne Y, Martins CIM, Stoeck T, Pawlowski J. Supervised machine learning outperforms taxonomy-based environmental DNA metabarcoding applied to biomonitoring. Molecular Ecology Resources. 2018 Nov;18(6):1381-91.

[6] Braikia H, Ben Hamida S, Rukoz M. Random Forest Classifier for Marine Biodiversity Analysis. In: 2024 International Conference on Intelligent Systems and Computer Vision (ISCV); 2024. p. 1-8.

[7] Van Dam H. On the use of measures of structure and diversity in applied diatom ecology. Nova Hedwigia. 1982;73:97-115.

[8] Thukral A. A review on measurement of Alpha diversity in biology. Agricultural Research Journal. 2017 01;54:1.

[9] Shannon CE. A mathematical theory of communication. The Bell System Technical Journal. 1948;27(3):379-423.

[10] Pielou EC. The measurement of diversity in different types of biological collections. Journal of Theoretical Biology. 1966;13:131-44. Available from: https://www.sciencedirect.com/science/article/pii/0022519366900130.

[11] SIMPSON EH. Measurement of Diversity. Nature. 1949 Apr;163(4148):688-8. Available from: https://doi.org/10.1038/163688a0.

[12] Hazra A, Gogtay N. Biostatistics Series Module 10: Brief Overview of Multivariate Methods. Indian Journal of Dermatology. 2017 Jul-Aug;62(4):358-66.

[13] Bartholomew DJ. Analysis and Interpretation of Multivariate Data. In: Peterson P, Baker E, McGaw B, editors. International Encyclopedia of Education (Third Edition). third edition ed. Oxford: Elsevier; 2010. p. 12-7. Available from: https://www.sciencedirect.com/science/article/pii/B9780080448947013038.

[14] ter Braak CJF. Canonical Correspondence Analysis: A New Eigenvector Technique for Multivariate Direct Gradient Analysis. Ecology. 1986;67(5):1167-79. Available from: https://esajournals.onlinelibrary.wiley.com/doi/abs/10.2307/1938672.

[15] Lundberg SM, Lee SI. A unified approach to interpreting model predictions. In: Proceedings of the 31st International Conference on Neural Information Processing Systems. NIPS'17. Red Hook, NY, USA: Curran Associates Inc.; 2017. p. 4768–4777.

[16] Breiman L. Random Forests. Mach Learn. 2001 Oct;45(1):5–32. Available from: https://doi.org/10.1023/A:1010933404324.

Developing machine-learning-based amyloidogenicity predictors with Cross-Beta DB

Valentin GONAY[1,2], Michael P. DUNNE[2], Javier CACERES[3], Andrey V. KAJAVA[1]

1 CRBM UMR 5237 CNRS Université Montpellier, 1919 Rte de Mende, 34293, Montpellier, France

2 PROTERA SAS, 176 avenue Charles de Gaulle, 92522, Neuilly-sur-Seine Cedex, France

3 Protera (GEAEnzymes SpA), Av. Santa Maria 2810, Oficina 302 Providencia, Santiago, Chile

Corresponding Author: mdunne@proterabio.com, Andrey.Kajava@crbm.cnrs.fr

Paper Reference: Gonay *et al.* (2025) Developing machine-learning-based amyloidogenicity predictors with Cross-Beta DB. *Alzheimer's Dement.* https://doi.org/10.1002/alz.14510

Keywords

Machine-learning, Amyloids, Prediction, Neurodegenerative diseases, Database

Abstract

The importance of protein amyloidogenesis, associated with various diseases and functional roles, has driven the creation of computational predictors of amyloidogenicity. The accuracy of these predictors, particularly those utilizing artificial intelligence technologies, heavily depends on the quality of the data. We built Cross-Beta DB, a database containing high-quality data on known cross-β amyloids formed under natural conditions. We used it to train and benchmark several machine-learning (ML) algorithms to predict amyloid-forming potential of proteins. We developed the Cross-Beta predictor using an Extra trees ML algorithm, which outperforms other amyloid predictors with the highest F1 score (0.852) and accuracy (0.844) compared to existing methods. The development of the Cross-Beta DB database and a new ML-based Cross-Beta predictor may enable the creation of personalized risk profiles for neurodegenerative diseases and other amyloidoses—especially as genome sequencing becomes more affordable.

Highlight

A significant number of neurodegenerative diseases remain without effective treatments. This is particularly true for conditions triggered by amyloid proteins, collectively known as amyloidoses. These disorders include, but are not limited to, Alzheimer's disease, Parkinson's disease, and type 2 diabetes. The class of protein structures responsible for these conditions is referred to as amyloids, which are characterized by their distinctive structure and aggregation behavior. While amyloid proteins in their normal functional state are typically highly disordered (meaning they lack a stable three-dimensional structure) they can adopt a stable cross-beta structure when forming amyloids. In this fibrillar arrangement, polypeptide chains assume beta-conformations and align perpendicularly to the fibril axis. This unique structural organization makes amyloid fibrils remarkably resistant to degradation, including by proteases and extreme temperatures. Notably, amyloids are also prone to polymorphism, meaning a single amyloid sequence can give rise to diverse structural forms, even under identical environmental conditions [1].

Advancements in machine learning have enabled the analysis and classification of large datasets. Supervised models, such as the Random Forest and Extra Trees classifiers, are particularly notable examples. These models can effectively handle moderate-sized datasets, unlike unsupervised learning methods like neural networks, which often require millions of data points to perform optimally. This ability to operate efficiently with smaller datasets makes supervised models especially valuable for studying small protein populations involved in amyloid formation.

In our study, we highlight the importance of data selection in training effective models. For the positive dataset, we selected a group of proteins that form cross-beta amyloids under physiological conditions, specifically in terms of pH, temperature, and concentration. Since no existing database covered this range of data, we created our own: Cross-Beta DB. For the negative dataset, we included intrinsically disordered regions (IDRs) and intrinsically disordered proteins (IDPs) known to remain soluble under the same conditions. These negative data were sourced from the DisProt database [2]. A crucial next step was the development of features based on protein sequences. We designed a set of features that incorporated amino acid composition, di-peptide composition, and predicted structural properties. To refine our model, we identified key features by analyzing their relative importance in prediction performance. By leveraging these features, we optimized various supervised classifiers and ultimately selected the model that demonstrated the best performance for this task.

To evaluate our predictor's performance against similar models, we selected six established predictors: ArchCandy2.0 [3], AMYPredFRL [4], Aggrescan [5], PASTA2.0 [6], AmyloGram [7], and Tango [8]. We assessed these predictors using the same data we employed to create our testing sets: 10 subgroups, each containing 13 positive samples (cross-beta-forming amyloids) and 13 negative samples (soluble IDRs or IDPs). The F1 score was calculated as the average result across these 10 subgroups and compared to the performance observed during our predictor's testing phase. Our model, Cross-Beta Predictor, achieved an F1 score of 0.852, outperforming the other predictive models for amyloidogenicity.

These results highlight the importance of selecting an appropriate methodology for effectively addressing classification problems. With a well-designed approach, machine learning techniques can be successfully applied even to small datasets. This methodology can also be extended to other types of protein aggregation, such as liquid-liquid phase separation, gelling proteins, or more specific cases like examining the impact of missense mutations on protein amyloidogenicity [9].

References

1. Tycko R. Amyloid Polymorphism: Structural Basis and Neurobiological Relevance. Neuron. [Internet] 2015 [cited 2025 Feb 18];86(3):632–45. Available from: https://doi.org/10.1016/j.neuron.2015.03.017

2. Aspromonte MC, Nugnes MV, Quaglia F, Bouharoua A, Tosatto SC, Piovesan D. DisProt in 2024: improving function annotation of intrinsically disordered proteins. Nucleic Acids Research. [Internet] 2024 [cited 2025 Mar 12];52(D1), D434-D441. Available from https://doi.org/10.1093/nar/gkad928

3. Ahmed AB, Znassi N, Château MT, Kajava AV. A structure-based approach to predict predisposition to amyloidosis. Alzheimer's & Dementia. [internet] 2015 [cited 2025 Feb 21];11(6):681-90. Available from https://doi.org/10.1016/j.jalz.2014.06.007

4. Charoenkwan P, Ahmed S, Nantasenamat C, Quinn JM, Moni MA, Lio' P, Shoombuatong W. AMYPred-FRL is a novel approach for accurate prediction of amyloid proteins by using feature representation learning. Scientific reports. [Internet] 2022 [cited 2025 Feb 21];12(1):7697. Available from https://doi.org/10.1038/s41598-022-11897-z

5. Conchillo-Solé O, de Groot NS, Avilés FX, Vendrell J, Daura X, Ventura S. AGGRESCAN: a server for the prediction and evaluation of" hot spots" of aggregation in polypeptides. BMC bioinformatics. [Internet] 2007 [cited 2025 Feb 21];8:1-7. Available from https://doi.org/10.1186/1471-2105-8-65

6. Walsh I, Seno F, Tosatto SC, Trovato A. PASTA 2.0: an improved server for protein aggregation prediction. Nucleic acids research. [Internet] 2014 [cited 2025 Feb 21];42(W1):W301-7. Available from https://doi.org/10.1093/nar/gku399

7. Burdukiewicz M, Sobczyk P, Rödiger S, Duda-Madej A, Mackiewicz P, Kotulska M. Amyloidogenic motifs revealed by n-gram analysis. Scientific reports. [Internet] 2017 [cited 2025 Feb 21];7(1):12961. Available from https://doi.org/10.1038/s41598-017-13210-9

8. Fernandez-Escamilla AM, Rousseau F, Schymkowitz J, Serrano L. Prediction of sequence-dependent and mutational effects on the aggregation of peptides and proteins. Nature biotechnology. [Internet] 2004 [cited 2025 Feb 21];22(10):1302-6. Available from https://doi.org/10.1038/nbt1012

9. Pyankov IA, Gonay V, Stepanov YA, Shestun P, Kostareva AA, Uspenskaya MV, Pyankov MG, Kajava AV. A computational approach to predict the effects of missense mutations on protein amyloidogenicity: A case study in hereditary transthyretin cardiomyopathy. Journal of Structural Biology. [Internet] 2025 [cited 2025 Mar 11];217(1):108176. Available from https://doi.org/10.1016/j.jsb.2025.108176

Joint Embedding-Classifier Learning for Interpretable Collaborative Filtering

Clémence REDA[1,2], Jill-Jênn VIE[3] and Olaf WOLKENHAUER[1,4,5]

1 Institute of Computer Science, University of Rostock, Ulmenstrasse 69, 18051, Rostock, Germany

2 BioComp, Institut de Biologie de l'ENS (IBENS), CNRS, 46 rue d'Ulm, 75005 Paris, France

3 Soda, Inria Saclay, 1 rue Honoré d'Estienne d'Orves, 91120, Palaiseau, France

4 Leibniz-Institute for Food Systems Biology, Lise-Meitner-strasse 34, 85354, Freising, Germany

5 Stellenbosch Institute of Advanced Study, Wallenberg Research Centre, 10 Marais Road, 7602, Stellenbosch, South Africa

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

Paper Reference: Réda *et al.* (2025) Joint embedding–classifier learning for interpretable collaborative filtering. *BMC Bioinformatics*.
http://dx.doi.org/10.1186/s12859-024-06026-8

Keywords
Drug repurposing, Interpretability, Embedding learning, Pathway enrichment.

Abstract

Background: Interpretability is a topical question in recommender systems, especially in healthcare applications. An interpretable classifier quantifies the importance of each input feature for the predicted item-user association in a non-ambiguous fashion. Results: We introduce the novel Joint Embedding Learning-classifier for improved Interpretability (JELI). By combining the training of a structured collaborative-filtering classifier and an embedding learning task, JELI predicts new user-item associations based on jointly learned item and user embeddings while providing feature-wise importance scores. Therefore, JELI flexibly allows the introduction of priors on the connections between users, items, and features. In particular, JELI simultaneously (a) learns feature, item, and user embeddings; (b) predicts new item-user associations; (c) provides importance scores for each feature.

Moreover, JELI instantiates a generic approach to training recommender systems by encoding generic graph-regularization constraints. Conclusions: First, we show that the joint training approach yields a gain in the predictive power of the downstream classifier. Second, JELI can recover feature-association dependencies. Finally, JELI induces a restriction in the number of parameters compared to baselines in synthetic and drug-repurposing data sets.

Highlight

This paper proposes a flexible and scalable approach to, first, incorporating prior biological knowledge as a graph (e.g., protein-protein interaction networks, or generic knowledge graphs such as PrimeKG [1]) to a semi-supervised classification task of drug repurposing; and, second, to provide interpretability on the classification scores by relating them to specific nodes in the knowledge graph. In particular, we show that traditional pathway enrichment analyses (e.g., Gene Set Enrichment Analysis [2]) can be applied after classification to connect disease-perturbed gene expression to relevant functional pathways in melanoma, using the inferred importance scores. We propose an extensive experimental study, as JELI is compared to three strong baselines in drug repurposing, on four open-source datasets and with seven types of prior knowledge graphs. Furthermore, once trained, the proposed model JELI can be used without further training on unseen drugs and diseases thanks to the structure of the newly introduced classifier model. Moreover, JELI can be applied for any matching task, for instance, connecting genes to diseases instead of drugs as done in [3]. Finally, the method is available as open-source code on PyPI and on GitHub [4].

References

1. Chandak, P., Huang, K. & Zitnik, M. Building a knowledge graph to enable precision medicine. Sci Data 10, 67 (2023). Available from: https://doi.org/10.1038/s41597-023-01960-3

2. Subramanian, A., Tamayo, P., Mootha, V.K., Mukherjee, S., Ebert, B.L., Gillette, M.A., Paulovich, A., Pomeroy, S.L., Golub, T.R., Lander, E.S. & Mesirov, J.P. Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles, Proc. Natl. Acad. Sci. U.S.A. 102, 43:15545-15550 (2005). Available from: https://doi.org/10.1073/pnas.0506580102

3. Chen, H. & Zhang, Z. A semi-supervised method for drug-target interaction prediction with consistency in networks. PLoS One, 7;8 5:e62975 (2013). Available from: https://doi.org/10.1371/journal.pone.0062975

4. Reda, C., Vie, J.J. & Wolkenhauer, O. RECeSS-EU-Project/JELI: JELI v1.0.2. Zenodo (2025). Available from: https://doi.org/10.5281/zenodo.14753619

Mini-Symposiums

Comment concilier nos activités en bioinformatique avec les limites planétaires ?

David BENABEN[1], Victoria BOURGEAIS[2], Aurélie BUGEAU[2], Olivier GAUWIN[2], Gael GUENNEBAUD[3], Sophie SCHBATH[4]

1 INRAE, Univ. Bordeaux, UMR BFP, 33882, Villenave d'Ornon, France

2 Univ. Bordeaux, CNRS, Bordeaux INP, LaBRI, UMR 5800, F-33400 Talence, France

3 Inria, Université de Bordeaux

4 Université Paris-Saclay, INRAE, MaIAGE, 78350, Jouy-en-Josas, France

Corresponding Author: david.benaben@inrae.fr, sophie.schbath@inrae.fr

Keywords

Crise environnementale, Responsabilité environnementale, Éthique de la recherche, Bioinformatique

Abstract

Ce mini-symposium transversal initie la discussion sur le sens de nos activités professionnelles au regard de la crise environnementale actuelle et de son évolution. Après un exposé introductif sur les **impacts environnementaux** dont le numérique, Stéphanie Mariette, chercheuse en génétique des populations à INRAE au sein de l'UMR Biogeco partage ses **réflexions** [1] quant à l'évolution récente de leur discipline, la génétique des populations, et à ses devenirs possibles, dans un scénario de résistance aux impératifs de la croissance, du big data et de l'innovation perpétuelle.

Dans un second temps, et avant une restitution en plénière, a lieu des **ateliers participatifs** en groupe plus restreint :

- A1 - **Usage du numérique** au sein du domaine de la bioinfo
- A2 - Prospective de nos domaines scientifiques sous **contraintes de ressources** (énergie, eau, processeurs, etc.)
- A3 - Intégrer les enjeux environnementaux dans la **conduite de la recherche** en bioinfo - une responsabilité éthique (avis du COMETS CNRS [2])
- A4 - **Cartographie des valeurs et attachements** (selon la philosophie des ateliers SEnS [3])

References

1. Sophie Gerber, Stéphanie Mariette. Les marqueurs du vivant : génétique et big data. 25 oct 2023; Disponible sur: https://www.terrestres.org/2023/10/25/les-marqueurs-du-vivant/

2. Comité d'éthique du CNRS (COMETS). Intégrer les enjeux environnementaux à la conduite de la recherche – Une responsabilité éthique [Internet]. 2022 déc. Report No.: 2022-43. Disponible sur: https://comite-ethique.cnrs.fr/avis-du-comets-integrer-les-enjeux-environnementaux-a-la-conduite-de-la-recherche-une-responsabilite-ethique/

3. Atelier SEnS [Internet]. Disponible sur: https://sens-gra.gitlabpages.inria.fr/atelier-impacts-recherche/

Mini-symposium JOBIM 2025: The Genomics of Biodiversity

Erwan CORRE[1], Alexandre LOUIS[2] and Hugues ROEST CROLLIUS[2]

1 Plateforme ABiMS/IFB,CNRS-Sorbonne Université, Station Biologique de Roscoff FR2424, France

2 Institut de Biologie de l'ENS (IBENS), CNRS UMR8197, 75005 Paris, France

Corresponding authors: corre@sb-roscoff.fr, alouis@bio.ens.psl.eu, hrc@bio.ens.psl.eu

Keywords

Biodiversity, Genomics, Evolution, Marine Biology, Infrastructures.

Description

The advent of large-scale sequencing initiatives, such as those coordinated by the **Earth BioGenome Project (EBP)**, is transforming the field of comparative genomics. Major challenges include integrating this unprecedented volume of data, managing its inherent complexity, and leveraging its vast taxonomic breadth to deepen our understanding of genome evolution and biodiversity.

Several pioneering projects exemplify these efforts. The **Vertebrate Genome Project (VGP)** is nearing completion of its first phase, delivering one genome per vertebrate order, totaling almost 600 high-quality assemblies, the **Darwin Tree of Life (DToL)** , which aims to sequence all eukaryotic species in the British Isles and Ireland, has produced almost 2,000 genome assemblies and the **European Reference Genome Atlas (ERGA)** brings together a diverse European research community to sequence continental Europe's biodiversity. In France, the **ATLASea program** is building capacity to sequence 4,500 marine species over the next seven years, with the aim of better understanding the evolution of marine species, the function of their genomes and the dynamics of marine ecosystems. To address the scientific and technical challenges posed by these transformative initiatives, the **BYTE-Sea project** within ATLASea proposes a mini-symposium on biodiversity genomics, with a specific focus on eukaryotic comparative genomics.

Invited speakers **Josefin Stiller** (University of Copenhagen), **Matthieu Muffato** (Welcome Sanger Institute)**, Yannis Nevers** (University of Strasbourg) and **Elise Parey** (University College London) will present their work on **Bioinformatics developments** needed to scale up methods to analyse thousands of genomes simultaneously and **applications** that embrace the complexity of multi-genome datasets to uncover novel insights into genome evolution, organisation, and function.

Mini-symposium : AI in Healthcare: From Fundamentals to the Clinic

Artificial Intelligence (AI) is transforming healthcare, from early diagnosis to personalized treatment. This mini-symposium explores the real-world impact of AI in clinical practice, with a focus on its integration into diagnostic and therapeutic workflows. By bridging fundamental research and clinical application, the event brings together researchers, clinicians, and AI experts to promote interdisciplinary exchange on both the promises and limitations of AI in medicine. Topics include methodological rigor, regulatory requirements, and the evolving role of healthcare professionals in an increasingly data-driven environment.

Jean-Marc Alliot is a Scientific Director of AI & Data in Toulouse University Hospital. He provides a critical historical overview of AI in medicine, highlighting both major advances and the importance of methodological soundness in clinical translation. **Title:** Artificial Intelligence and Medicine: A “Success Story”

Daniel Racoceanu is Professor at Sorbonne University & Paris Brain Institute (ICM) and presents two use cases: *PhagoStat*, an interpretable deep learning pipeline for quantifying phagocytosis in neurodegeneration studies; *Virtual Staining*, a generative AI system replacing traditional chemical stains with multi-stain prediction from H&E slides. Both tools are open-source and aim to make AI robust, interpretable, and sustainable. **Title:** Explainable AI in Biomedical Imaging – From PhagoStat to Virtual Staining

Simon Cabello-Aguilar is Bioinformatics Engineer in Montpellier University Hospital and presents a validated clinical pipeline for identifying MET amplification in NSCLC patients. Based on data from 1,932 patients, the tool supports therapeutic stratification and trial inclusion decisions. **Title:** AI-assisted NGS Detection of MET Amplification in Lung Cancer

Acknowledgements

We thank our sponsors for their support:

Organising Committee

Delphine Potier (CRCM, SFBI) - Elodie Darbo (BRIC/LaBRI, SFBI) - Laetitia Bourgeade (CHU Bordeaux, BioinfoDiag) - Charles Van Goethem (CHU Montpellier, BioinfoDiag)

Methods for Interfacing with Graphs of Genomic Sequences: novel Pangenome Paradigms

Sèverine BERARD[1], Guillaume GAUTREAU[2], Claire LEMAITRE[3], Jean MONLONG[4], François SABOT[5], Camille MARCHET[6], Benjamin LINARD[7]

1 ISEM, Univ Montpellier, CNRS, IRD, Montpellier, France

2 Université Paris-Saclay, INRAE, MaIAGE, 78350, Jouy-en-Josas, France

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

4 IRSD - Digestive Health Research Institute, University of Toulouse, INSERM, INRAE, ENVT, UPS, Toulouse, France

5 DIADE University of Montpellier Cirad IRD, 911 avenue Agropolis, 34394, Montpellier Cedex 5, France

6 Univ. Lille, CNRS, Centrale Lille, UMR 9189 CRIStAL, F-59000 Lille, France

7 Univ. Toulouse, Unité MIAT, INRAE Occitanie-Toulouse, 24 Chemin de Borde Rouge , 31320 Auzeville-Tolosane FRANCE

Corresponding Author: benjamin.linard@inrae.fr

Keywords

Pangenome graph, genetic diversity, structural variations

Abstract

The development of genome sequencing offered many opportunities to explore the function and evolution of species. A standard emerged in the 2000s: omics studies, in all their diversity, were contextualized with a reference genome, allowing the identification of variations held by new sequences relative to this shared reference. The limits of this approach appeared rapidly in procaryote species and led to the concept of pangenomes. With the recent availability of highquality, telomere-to-telomere, genome sequencing, the concept of pangenome can now be applied even to the most complex eukaryotes. More generally, new models such as pangenome graphs or pangenome databases allow to compare new sequences to the whole diversity of genome variation known for a species or a species complex. This approach reduces analysis biases, and is promising for extended exploration of complex structural variations (SV), a family of variations that remains poorly explored when compared to SNPs.

The paradigm change of “pangenomes as the new reference” opens a range of new challenges in our community. Many bioinformatic analyses are oriented towards a reference and not pangenome-based. Since a few years, tools are nevertheless diversifying and pangenome research and engineering is spreading in many fields, from microbiology, health,

agronomy to environmental sciences.

This mini-symposium aims to gather the research community interested in the many methodological and computational challenges related to pangenome models and in particular graph-based models. Via short talks and a contribution from an invited speaker, leader in pangenome graph developments, we will discuss some identified problems such as pangenome construction scalability, visualisation, and exploitation. We will also run an open table discussion where beginners & specialists will discuss their own problematic and expectations for pangenomes approaches, and how they could help to explore further their scientific question.

Program overview

The objective of this mini-symposium is to bring together both new and experienced researchers engaged in, or simply curious about computational pangenomics, with some emphasis on graph-based models.

INVITED SPEAKER: JANA EBLER, Institute for Medical Biometry and Bioinformatics, Heinrich Heine University Düsseldorf, Germany

Pangenome-based genome inference. Typical analysis workflows map reads to a reference genome in order to genotype genetic variants. Generating such alignments introduces reference biases and comes with substantial computational burden. In contrast, recent k-mer based genotypers are fast, but struggle in repetitive or duplicated genomic regions. We introduced a new algorithm, PanGenie, that leverages a haplotype-resolved pangenome reference in conjunction with k-mer counts from short-read sequencing data to genotype a wide spectrum of genetic variation – a process we refer to as genome inference. Improvements are especially pronounced for structural variants (SVs) and variants in repetitive regions. We studied SVs across large cohorts sequenced with short-reads, using pangenome graphs generated by the HGSVC and HPRC consortia, which enables the inclusion of these classes of variants in genome-wide association studies.

FLASH TALKS: Highlights of JOBIM posters related to pangenomic approaches.

ROUND TABLE: Do we need a pangenome graph? What is a good pangenome? A debate moderated by a panel of developers and users. Everyone is welcome, from experienced developers to researchers wondering if these approaches may be beneficial to their research.

GDR Groupement de recherche
BIMMM Bio-Informatique Moléculaire : Modélisation et Méthodologie