JOBIM 2026

30 JUIN>3 JUIL

Strasbourg

Proceedings

Preface

Dear Participants and Colleagues,

We are very pleased to present the proceedings of the 2026 edition of JOBIM, hosted this year in Strasbourg, which brought together 500 participants on-site and nearly 100 joining online.

We are delighted to present a high-quality scientific program, drawn from a total of **265 submissions**. Following a rigorous peer-review process, the final selection features **67 oral presentations** and **185 posters**, organized as follows:

- **Original Research:** 33 talks and 20 posters (out of 56 submissions),
- **Previously Published Work:** 6 talks (out of 8 submissions),
- **Full Proceedings Papers:** 3 talks and 2 posters (out of 6 submissions),
- **Full Proceedings Papers and PCI:** 2 talks and 1 poster (out of 3 submissions),
- **Platforms & Services:** 1 talk and 3 posters (out of 4 submissions),
- **Demos:** 11 oral demonstrations (out of 11 submissions),
- **Posters:** 163 posters (out of 177 submissions).

We hope the choices we made reflect the diversity, innovation, and collaborative spirit of the JOBIM 2026 community.

In addition to these tracks, we are pleased to highlight the four mini-symposia proposals selected by the Program Committee, which will enrich Thursday afternoon with an exciting and diverse program covering:

- Spatial Biology: principles, tools and applications for translational research in oncology
- International bioinformatics connection spotlighting sequencing technologies applied to One Health
- Is the structural annotation of genes in eukaryotic genomes still a challenge?
- Orchestrating data flows throughout their lifecycle

We are also particularly honored to welcome six distinguished keynote speakers whose insights will greatly enriched this edition of JOBIM: Hugues Roest-Crollius (ENS Paris), Laura Cantini (Pasteur Paris), Christophe Dessimoz (SIB, Lausanne), Daniel Jost (LBMC, ENS Lyon), Sarah Cohen-Boulakia (LISN, Paris Saclay) and Judith Zaugg (University of Basel).

We warmly thank all reviewers, authors, and contributors who made this scientific program possible. We would also like to express our sincere thanks to our institutional partners, SFBI, IFB and GdR BIMMM, and all the sponsors. Their support and commitment have been essential to the success of this event. Finally, we thank everyone who contributed to making JOBIM 2026 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.

Isabel Alves, Anaïs Bardet, Claudine Mayer

The Program Committee

Chaired by Isabel Alves, Anaïs Bardet, and Claudine Mayer

- Sophie Abby
- Isabel Alves
- Gwenaëlle André
- Benoit Ballester
- Anaïs Bardet
- Anaïs Baudot (GDR BIMMM)
- Séverine Bérard
- Camille Berthelot
- Audrey Bihouée (SFBI)
- Yuna Blum
- Laura Cantini
- Hélène Chiapello
- Erwan Corre (IFB)
- Olivier Dameron
- Elodie Darbo (SFBI)
- Sandra Derozier (SFBI)
- Sarah Djebali
- Damien Eveillard
- Anna-Sophie Fiston-Lavier
- Christine Gaspin (IFB)
- Vincent Lacroix
- Charles Lecellier
- Claire Lemaitre
- Emmanuelle Lerat
- Camille Marchet
- Mahendra Mariadassou
- Florian Massip
- Claudine Mayer
- Jean Monlong
- Anamaria Necsulea
- Yannis Nevers
- Bernard Offmann
- Loïc Paulevé
- Olivier Poch
- Yann Ponty
- Delphine Potier
- Fanny Pouyet
- Magali Richard
- Eric Rivals
- Céline Scornavacca
- Nelle Varoquaux
- Nathalie Vialaneix
- Matthias Zytnicki

Table of Contents

Keynotes

Oral Presentations

Oral Full Proceedings

Demos

Posters

Keynotes

Deciphering the structure-function relationship of chromatin: from experiments to modeling and back

Keynote

Daniel JOST [1]

1. *Laboratoire de Biologie et Modélisation de la Cellule, École Normale Supérieure de Lyon, CNRS, UMR 5239, Inserm, U1293, Université Claude Bernard Lyon 1*

Abstract

Inside cells, DNA is packed into a polymer-like structure called chromatin. Characterizing how chromatin self-organizes is one of the major challenges faced in recent years by biology. During the last decade, thanks to the development of advanced experimental techniques coupled to biophysical modeling, major progresses have been realized in our understanding of the multi-scale chromosome organization. An increasing number of evidences has suggested that the spatio-temporal organization of the genome play a decisive role in the regulation of gene expression and in diseases. In this presentation, I will first briefly introduced the field of 3D Genomics and its current challenges, and then I will illustrate how polymer physics and numerical simulations were instrumental to better understand the mechanisms driving the complex spatial organization of chromosomes inside nuclei. In particular, I will focus on the role of the replication machinery in shaping chromatin architecture during the S-phase.

URL

D'Asaro, D., Arbona, JM., Piveteau, V. *et al.* Genome-wide modeling of DNA replication in space and time confirms the emergence of replication specific patterns in vivo in eukaryotes. *Genome Biol* **26**, 431 (2025). https://doi.org/10.1186/s13059-025-03872-4

Multi-modal learning for single-cell data integration

Keynote

Laura Cantini [1]

1. *Institut Pasteur*

Abstract

Single-cell RNA sequencing (scRNAseq) is revolutionizing biology and medicine. The possibility to assess cellular heterogeneity at a previously inaccessible resolution, has profoundly impacted our understanding of development, of the immune system functioning and of many diseases. While scRNAseq is now mature, the single-cell technological development has shifted to other large-scale quantitative measurements, a.k.a. 'omics', and even spatial positioning.

Each single-cell omics presents intrinsic limitations and provides a different and complementary information on the same cell. The current main challenge in computational biology is to design appropriate methods to integrate this wealth of information and translate it into actionable biological knowledge.

In this talk, I will discuss three main computational directions currently explored in my team: (i) dimensionality reduction to study cellular heterogeneity simultaneously from multiple omics; (ii) gene network inference to integrate a large range of interactionsbetween the features of various omics and isolate the regulators underlying cellular heterogeneity and (iii) spatially-informed trajectory inference methods to reconstruct the spatiotemporal landscape underlying cell dynamics.

Open biological databases as strategic infrastructure: from research to competitiveness and sovereignty

Keynote

Christophe Dessimoz [1]

1. *SIB Swiss Institute of Bioinformatics*

Abstract

Open biodata resources are critical to modern life science, yet many of them remain chronically underfunded and increasingly vulnerable. This is the paradox at the centre of the talk. I will outline the role of open biodata resources in research, innovation, and major societal challenges such as public health, food security, and biodiversity conservation. I will also address their role in helping Europe advance its competitiveness and sovereignty priorities, especially in the age of AI. Finally, I will outline practical paths forward to secure these resources as strategic infrastructure for science, policy, and society.

URL

https://www.nature.com/articles/s41597-026-06690-w https://www.nature.com/articles/s41597-024-03099-1

Reproducibility by Design in Bioinformatics: Research challenges and opportunities

Keynote

Sarah Cohen-Boulakia [1]

1. *Université Paris-Saclay, CNRS, Laboratoire Interdisciplinaire des Sciences du Numérique*

Abstract

Reproducibility has become a central concern in computational biology, yet it remains difficult to achieve in practice. In this keynote, we examine reproducibility not as an afterthought but as a design principle that should guide how bioinformatics research is conducted, shared, and reused.The bioinformatics landscape has shifted considerably over the past decade, with increasingly code-centric workflow systems: Snakemake, Nextflow and their ecosystems, becoming the new standard. This transition brings new opportunities for reproducibility, but also new challenges the field is only beginning to address.We present complementary efforts spanning large-scale studies of workflow sharing and reuse, interactive visualization approaches, and reconciliation methods bridging scientific articles and executable code. Together, these contributions argue that reproducibility is not a single problem but a constellation of challenges and that the most promising opportunities lie precisely at the intersection of data science, visualization, and natural language processing.

The evolution of 568 vertebrate genomes during 400 million years

Keynote

François Giudicelli [1], ***Alexandra Louis*** [1], ***vgp phasel*** [2], ***Hugues Roest Crollius*** [3]

1. *Institut de Biologie de l'ENS (IBENS), Département de biologie, École normale supérieure, CNRS, INSERM, Université PSL,* ***2.*** *vertebrategenomesproject,* ***3.*** *Equipe DYOGEN, Institut de Biologie de l'ENS (IBENS), Département de biologie, École normale supérieure, CNRS (UMR8197), INSERM (U1024), Université PSL, 75005 Paris, France*

Abstract

Vertebrates comprise approximately 70,000 living species that evolved from a common ancestor some 500 million years ago. Over this immense timespan, vertebrate genomes have diversified through chromosome rearrangements, gene gains and losses, and changes in nucleotide composition, while retaining traces of their shared ancestry. Reconstructing the genomic architecture of early vertebrates is therefore essential to understand when key features of modern genomes first emerged, including microchromosomes, conserved syntenic blocks, and large-scale compositional heterogeneity. Yet such reconstructions have long remained elusive because they require dense phylogenetic sampling and high-quality genome assemblies. The completion of Phase I of the Vertebrate Genome Project (VGP) marks a turning point. Over the past seven years, the consortium has generated 581 chromosome-level reference genomes spanning approximately 85% of vertebrate orders, providing an unprecedented resource for evolutionary genomics. I will first present the structure of the VGP consortium and the genomic resources it has produced. I will then describe AGORA 2.0, a new version of our ancestral genome reconstruction framework, which we applied to infer the gene order and chromosomal organisation across 565 ancestral genomes. These reconstructions reveal that major vertebrate lineages have followed very distinct patterns of genome evolution since their common ancestors. While sauropsids have remained remarkably stable, mammals have undergone numerous inter-chromosomal fusions/fissions, while ray-finned fish genomes were subject to intensive intra-chromosomal rearrangements. They uncover the deep evolutionary origins of GC-rich microchromosomes, whose signatures remain detectable in present-day genomes. They further show that bird "dot chromosomes" follow a distinctive evolutionary trajectory characterised by elevated rearrangement rates and recurrent clustering and fusion events during mammalian and crocodilian evolution. Beyond these biological insights, the project delivers a freely accessible resource for comparative genomics, available through a dedicated Genomicus-VGP portal (https://www.genomicus.bio.ens.psl.eu/genomicus-vgp-02.01/), enabling researchers to explore vertebrate genome evolution across 500 million years of evolutionary history.

Understanding disease mechanisms through the lens of gene regulation at single-cell and spatial resolution

Keynote

Judith Zaugg [1]

1. *Université de Bâle*

Abstract

Individuals differ not only in their genome but in how their cells respond to signals, interact with their environment, and maintain or lose functional states over time. In this lecture I will argue that this variation, genetic and non-genetic alike, is a powerful resource for learning disease mechanisms, if we can interpret it through the lens of gene regulation.

I will present computational frameworks developed in my group to infer gene regulatory networks from single-cell epigenomic and transcriptomic data, and show how applying them to variation across individuals and cell states reveals the regulatory logic of haematopoiesis. Focusing on the bone marrow niche, I will illustrate how stromal, immune, and stem cells mutually influence each other, and how perturbation of this system, including acute myeloid leukaemia, can be traced back to specific regulatory programs. I will close with examples of how spatial genomics is beginning to reveal the tissue-scale organisation underlying these interactions, and reflect on the computational challenges that remain in moving from static snapshots to a dynamic, predictive understanding of cell states in disease.

Oral Presentations

A consensus-driven framework for building and sharing single-cell atlases applied to pancreatic ductal adenocarcinoma

Oral Presentation

Lucie Lamothe [1], ***Polina Arsenteva*** [2], ***Franck Picard*** [2], ***Yasmina Kermezli*** [1], ***Magali Richard*** [3], ***Yuna Blum*** [4]

1. *Université Grenoble Alpes,* ***2.*** *LBMC CNRS UMR 5239, Ecole Normale Supérieure de Lyon,* ***3.*** *CNRS,* ***4.*** *IGDR*

Abstract

Pancreatic ductal adenocarcinoma (PDAC) is a highly aggressive and invasive tumoral lesion affecting the pancreas. Molecular analysis and classification based on gene expression landscapes are complicated by the intrinsic heterogeneity of PDAC tumors. Like all solid cancers, PDAC consists of a 'tumor mass' (predominantly epithelial cells) which is surrounded by a microenvironment composed of stromal (fibroblasts, pericytes, endothelial) and immune cells, which provide support, nutrients and sometimes resistance or metastatic potential to neoplastic cells. Precise quantification of this tumor heterogeneity is of utmost importance, as these multiple components are key factors in explaining tumor progression and response to therapy.

A promising approach to accurately quantify cell type heterogeneity in PDAC relies on the recent emergence of bulk and spatial deconvolution algorithms based on single-cell reference profiles [1]. One of the main limitations of these approaches is the accuracy of the single-cell-based profiles, which can strongly impair the quantification and the biological interpretation of the inferred tumor composition [2]. To overcome these difficulties, we built an integrative set of PDAC cell-type-specific gene markers, based on a dedicated pre-established gene markers curation, and subsequent analysis and annotation of PDAC recent single-cell RNA-seq datasets within a consensus-driven framework.

Using these finely annotated datasets we are running systemic identification of integrative new PDAC cell-types-specific gene markers. We then intend to use these annotated datasets and associated newly uncovered markers to revise our current understanding of cell-type heterogeneity in PDAC, using single-cell based bulk deconvolution approaches and spatial transcriptomic analysis.

In this talk, we focus on our annotation approach consensus based and perspectives offered by such annotated datasets.

References

[1] Francisco Avila Cobos et al. Genome Biology, 2023.

[2] Geng Chen et al. Frontiers in Genetics, 2019.

URL

/

A Probabilistic Framework for Clonal Reconstruction in chronic lymphocytic leukemia (CLL)

Oral Presentation

Vidhi Chhillar [1], ***Ulysse Herbach*** [1], ***Coralie Fritsch*** [1], ***Nicolas Champagnat*** [1]

1. *University of Lorraine*

Abstract

Intra-tumor heterogeneity is a significant challenge in targeted cancer therapy, particularly in chronic lymphocytic leukemia (CLL). Clonal evolution, driven by V(D)J recombination and somatic hypermutation, generates heterogeneous subclonal populations that can confer treatment resistance, making disease progression difficult to predict and treat. Phylogenetic reconstruction is key to understanding this process. However, existing methods such as ViCloD and B-SCITE return a single deterministic phylogeny, failing to capture uncertainty inherent in sequencing data and the biological ambiguity of clonal boundaries.

We present a probabilistic framework for clonal reconstruction based on the weighted tree distribution over trees. Rather than inferring a single tree, our approach defines an EM model with a conjugate prior that yields a tractable posterior distribution over trees. From this, we derive a phylogenetic graph summarising the ensemble of plausible evolutionary trajectories, explicitly encoding topological uncertainty arising from biological variability and sequencing noise.

To refine the inferred clonal structure, we introduce a Variational EM algorithm that iteratively infers an optimal number of latent, unobserved intermediate clones. At each E-step, the posterior over latent clones is approximated via mean-field variational inference; the M-step updates the graph structure to maximise the ELBO. By explicitly modelling missing evolutionary intermediates, the Variational EM resolves topological ambiguity that the standard EM cannot, yielding a more complete and accurate reconstruction of clonal dynamics.

While the discussion above focuses on a particular evolution model, the underlying framework is more general and can readily incorporate other evolution models such as JC69, GTR etc. whenever required.

A Statistical Workflow Combining Full-Scan and Targeted Analyses for Identifying Candidate Volatile Compounds from SIFT-MS Data

Oral Presentation

Axel Mercier [1]

1. *Multitel*

Abstract

The gut microbiota is involved in numerous pathological conditions, and identifying non-invasive biomarkers of its metabolic activity remains a major challenge. The volatolome — the set of volatile compounds present in exhaled breath — offers a promising avenue to address this issue.

We propose a statistical workflow combining full-scan and targeted analysis using Selected Ion Flow Tube Mass Spectrometry (SIFT-MS) to identify candidate volatile compounds from breath samples. Full-scan acquisition is first performed across all samples, generating an ion signal matrix (precursor ion, m/z). Dimensionality reduction via UMAP enables visualization of the overall data structure and identification of time points with the most pronounced group separation. Statistical tests are then applied to the signal matrix, and discriminant signals are used to infer a list of candidate volatile compounds from the SIFT-MS database. These candidates are subsequently evaluated through targeted SIFT-MS analysis, allowing concentration estimates and a second round of statistical testing to refine the final list before biological validation.

Applied to a pilot dataset of two groups of 14 paired participants, the workflow identified several discriminant ionic signals and candidate volatile compounds consistent with existing literature. Comparing results from both analytical stages allows assessment of workflow concordance prior to biological validation.

This work provides a reproducible methodological framework for the statistical exploration of SIFT-MS data, aimed at facilitating the identification of non-invasive respiratory biomarkers.

Advancing Neisseria Gonorrhoeae Surveillance through Long-Read Sequencing, Pangenome Graphs, and Mass Spectrometry.

Oral Presentation

Christian Blumenscheit [1], ***Kathleen Klaper*** [2], ***Martin Hölzer*** [1], ***Dagmar Heuer*** [2], ***Hugues Richard*** [3]

1. *Genome Competence Center (MF1), Robert Koch Institute, Nordufer 20, 13353,* ***2.*** *Sexually transmitted bacterial pathogens (FG18) , Robert Koch Institute, Nordufer 20, 13353,* ***3.*** *Genome Competence Center (MF1), Robert Koch Institute, Nordufer 20, 13353 Berlin, Germany*

Abstract

Genomic sequencing is an important component of bacterial surveillance: beyond strain typing and profiling of Antimicrobial resistance (AMR), it improves contact tracing during outbreak analysis. Traditional methods using short-read sequencing have until now proven effective, but are inherently limited when considering complex genomic structures resulting from recombination.

We conducted a pilot study on 114 Neisseria gonorrhoeae (NGO) isolates from the German surveillance program and resequenced them using nanopore long reads. NGO is a sexually transmitted pathogen with increased AMR, making it a significant global health concern. Our results demonstrated high-resolution assembly: 108/114 isolates assembled into one single contig, and the median error rate was below 0.002% (after Illumina polishing). This enhanced resolution first enabled us to characterize strains at a finer grain: the mutations in the mtR promoter regions -associated with Azithromycin resistance, can be pinpointed at a bp resolution. We also identified a novel clade associated with mild cefixime resistance. This clade showed a higher mutation rate than ~800 sequenced isolates in Europe during the same period, suggesting it is under stronger evolutionary pressure.

We further integrated the strain collection in a gene-based pangenome graph using the Ppanggolin Software. The graph representation integrates contextual information beyond traditional presence/absence lists and enabled us to identify a 20 kb insert specific to a clade. We further developed an automated statistical technique to detect enriched gene sequences (as paths in the graph) across strain collections, revealing distinct functional profiles associated with resistance.

We then explored the integration of proteomic data by using mass spectrometry (timsTOF). Preliminary results show that peptide analysis can effectively distinguish clades, offering a rapid, cost-effective typing approach with the potential to transform bacterial surveillance and research. Our combined genomic and proteomic strategy, leveraging pangenome graph representations, represents a significant advancement in bacterial genomics.

Adversarial Domain Adaptation Enables Knowledge Transfer Across Heterogeneous RNA-Seq Datasets

Oral Presentation

Kevin Dradjat [1], ***Massinissa Hamidi*** [1], ***Blaise Hanczar*** [1]

1. *Université d'Evry Paris-Saclay*

Abstract

Accurate phenotype prediction from RNA sequencing (RNA-seq) data is essential for diagnosis, biomarker discovery, and personalized medicine. Deep learning models have demonstrated strong potential to outperform classical machine learning approaches, but their performance relies on large, well-annotated datasets. In transcriptomics, such datasets are frequently limited, leading to over-fitting and poor generalization. Knowledge transfer from larger, more general datasets can alleviate this issue. However, transferring information across RNA-seq datasets remains challenging due to heterogeneous preprocessing pipelines and differences in target phenotypes.

In this study, we propose a deep learning-based domain adaptation framework that enables effective knowledge transfer from a large general dataset to a smaller one for cancer and tissue type classification. The method learns a domain-invariant latent space by jointly optimizing classification and domain alignment objectives. To ensure stable training and robustness in data-scarce scenarios, the framework is trained with an adversarial approach with appropriate regularization. Both supervised and unsupervised approach variants are explored, leveraging labeled or unlabeled target samples.

The framework is evaluated on three large-scale transcriptomic datasets (TCGA, ARCHS4, GTEx) to assess its ability to transfer knowledge across cohorts. Experimental results demonstrate consistent improvements in cancer and tissue type classification accuracy compared to non-adaptive baselines, particularly in low-data scenarios.

Overall, this work highlights domain adaptation as a powerful strategy for data-efficient knowledge transfer in transcriptomics, enabling robust phenotype prediction under constrained data conditions.

URL

https://doi.org/10.48550/arXiv.2603.08062

AlignMarkers: a pipeline for accurate context sequence-based placing of molecular markers across genomes and assemblies

Oral Presentation

Camille Auneau [1], ***Baptiste Imbert*** [2], ***Mathieu Zemihi*** [2], ***Gregoire Aubert*** [2], ***Clement Lavaud*** [3], ***Nadim Tayeh*** [2], ***Marie-Laure Pilet-Nayel*** [3], ***Jonathan Kreplak*** [2]

1. *Université Bourgogne Europe, Institut Agro Dijon, INRAE, Agroécologie & IGEPP, INRAE, Institut Agro, Univ Rennes,* ***2.*** *Université Bourgogne Europe, Institut Agro Dijon, INRAE, Agroécologie,* ***3.*** *IGEPP, INRAE, Institut Agro, Univ Rennes*

Abstract

Advances in sequencing have greatly improved genome assembly quality while reducing costs, producing multiple versions for the same species. This creates challenges for reusing previously identified molecular markers (SNPs and indels), whose precise positioning is crucial for population genetics, trait mapping, and marker-assisted selection.

To address these issues, we developed AlignMarkers, a bioinformatic pipeline that accurately remaps molecular markers across genome assemblies using sequence alignment, originally designed for multi-species genotyping array design. It supports multiple input formats (VCF, BED, FASTA, CSV) and is optimized for sequences of 100 base pairs or more. It can work on short markers with only their context sequences or positionnal markers. In the later case, sequences are extracted from reference genomes with a chosen context length and aligned to target assemblies using Minimap2. Stringent filtering ensures correct alignments, verifies variants, and identifies multimapping events. Built with Nextflow and Python, AlignMarkers integrates Minimap2, Samtools, and Bedtools and produces reports and visualizations for easy interpretation. Its design efficiently handles large, heterogeneous marker datasets and provides a robust framework for cross-genome remapping.

To evaluate the pipeline and assess the impact of context sequence length on multimapping, 80,000 markers were randomly selected across diverse genomic regions and realigned to the same genome. All expected positions were recovered, including repeat-rich regions, and multimapping decreased as context length increased. AlignMarkers was applied to several real datasets, including pea (Pisum sativum L.) and soybean (Glycine max). For the GenoPea 13.2K SNP set, 96.7% of markers were accurately transferred onto the Cameor V2 pea genome, 88.6% of pea transcriptome SNPs were correctly positioned. For soybean, 99.2% of SNP and indel markers were successfully remapped. In all cases, markers that failed were due to multimapping.

By providing precise marker remapping and efficient dataset management, AlignMarkers enables diverse genomic applications, including genotyping array development.

URL

https://forge.inrae.fr/geapsi/pipeline/specifics_alignmarkers

Backtrack-free network propagation with in-degree normalization

Oral Presentation

Jędrzej Kubica [1], ***Dariusz Plewczynski*** [2], ***Sébastien Déjean*** [3], ***Nicolas Thierry-Mieg*** [1]

1. *Univ. Grenoble Alpes, CNRS, UMR 5525, BCM, TIMC, 38000, Grenoble, France,* ***2.*** *Centre of New Technologies, University of Warsaw, S. Banacha 2c, 02-097, Warsaw, Poland,* ***3.*** *Institut de Mathématiques de Toulouse, UMR5219, CNRS, UPS, Université de Toulouse*

Abstract

Network propagation is a promising strategy to identify new candidate genes underlying diseases and other phenotypic traits. By propagating a "signal" from seeds (e.g. known disease genes) through a network, for example a protein-protein interaction network (or interactome), it integrates network topology with current knowledge about disease-causing genes. Other genes that collectively receive most of the signal are likely to be associated with the disease as well. However, existing network propagation algorithms are biased toward highly connected protein "hubs". In random walk-type approaches, such proteins typically receive inflated scores and high positions in new candidate rankings.

Here, we present Guilt-by-association (GBA) centrality, a novel network propagation algorithm designed to avoid the common pitfalls and biases associated with hubs. We tested GBA centrality using a human interactome across four phenotypes, and compared it with two state-of-the-art network propagation methods: Random Walk with Restart (RWR) and NetCore. For each phenotype, we assessed whether GBA centrality could recover known phenotype-associated genes and whether new candidate genes were enriched in relevant tissues. Depending on the phenotype, GBA centrality outperformed or matched RWR and NetCore. Genes with intermediate or low degrees are of high interest in disease gene prioritization. GBA centrality performed better for left-out genes with intermediate or low degrees within the interactome, while RWR and NetCore performed better for left-out genes with high degrees. This demonstrated that the strategy employed in GBA centrality successfully counteracted the bias toward high-degree nodes that affects existing network propagation algorithms. In addition, we demonstrated that the high-scoring genes in GBA centrality were strongly enriched in the expected tissues. Although the high-scoring genes from RWR and NetCore were also somewhat enriched, this enrichment was significantly weaker.

In conclusion, GBA centrality is a powerful network propagation method and constitutes a strong alternative to existing network propagation algorithms.

URL

https://github.com/jedrzejkubica/GBA-centrality

BeeRNA: tertiary structure-based RNA inverse folding using Artificial Bee Colony

Oral Presentation

Mehyar MLAWEH [1], ***Tristan Cazenave*** [1], ***Inès Alaya*** [1]

1. *LAMSADE, Université Paris Dauphine*

Abstract

The Ribonucleic Acid (RNA) inverse folding problem, designing nucleotide sequences that fold into specific tertiary structures, is a fundamental computational biology problem with important applications in synthetic biology and bioengineering. The design of complex three-dimensional RNA architectures remains computationally demanding and mostly unresolved, as most existing approaches focus on secondary structures. In order to address tertiary RNA inverse folding, we present BeeRNA, a bio-inspired method that employs the Artificial Bee Colony (ABC) optimization algorithm. Our approach combines base-pair distance filtering with RMSD-based structural assessment using RhoFold for structure prediction, resulting in a two-stage fitness evaluation strategy. To guarantee biologically plausible sequences with balanced GC content, the algorithm takes thermodynamic constraints and adaptive mutation rates into consideration. In this work, we focus primarily on short and medium-length RNAs (< 100 nucleotides), a biologically significant regime that includes microRNAs (miRNAs), aptamers, and ribozymes, where BeeRNA achieves high structural fidelity with practical CPU runtimes. The lightweight, training-free implementation will be publicly released for reproducibility, offering a promising bio-inspired approach for RNA design in therapeutics and biotechnology.

URL

https://arxiv.org/abs/2511.21781

Circulating DNA reveals nucleosome occupancy patterns that are associated with nucleosome-DNA affinity and are affected in cancer

Oral Presentation

Marianne RICHAUD [1], ***Ekaterina Pisareva*** [1], ***Alain Thierry*** [1], ***Jacques Colinge*** [1]

1. *Institut de recherche en cancérologie de Montpellier*

Abstract

The study of cell-free circulating DNA (cirDNA) fragments (fragmentomics) from liquid biopsies has received increasing attention. By constructing an atlas of these well-positioned nucleosomes, which we called WPNA, we found that their occupancy was associated with histone-DNA affinity, as evidenced by codon usage bias and differences in cirDNA fragment sizes. Moreover, WPNA nucleosome occupancy was different in healthy and cancer samples, thus allowing developing a high performance machine learning approach for cancer detection (specificity and sensitivity >0.95 for seven cancer types). Cancer influenced WPNA nucleosome occupancy in a global manner, although distinct cancer types retained specific features. WPNA nucleosome occupancy at transcription factor binding sites revealed shared, pan-cancer regulation of transcriptional programs involved in hematopoietic cell differentiation and neutrophil biology, the main cirDNA sources. This work provides new fundamental insights into cirDNA and DNA sequence using cirDNA as a physical readout. It also bares translational significance by disclosing a new high-performance strategy for cancer detection from liquid biopsies.

URL

https://www.biorxiv.org/content/biorxiv/early/2025/10/09/2025.10.08.681110.full.pdf

Closing the sampling-scoring gap: a MassiveFold study in CASP16

Oral Presentation

Nessim Raouraoua [1], ***Thomas Binet*** [2], ***Marc Lensink*** [1], ***Guillaume Brysbaert*** [1]

1. *Univ. Lille, CNRS UMR 8576-UGSF-Unité de Glycobiologie Structurale et Fonctionnelle, 59000 Lille, France,* ***2.*** *US 41 – UAR 2014 – PLBS, University of Lille, CNRS, Inserm, CHU Lille, Institut Pasteur of Lille, 59000 Lille, France*

Abstract

The CASP15-CAPRI experiment demonstrated that massive-scale sampling, combined with diversity-inducing parameters (AFsample by Björn Wallner), significantly improves protein structure prediction. However, the original AlphaFold2 architecture presents a major bottleneck for large-scale sampling by being limited to single-process calculations. To address these computational limitations, we collaborated with Björn Wallner to create MassiveFold [1], a tool that enables massive sampling by parallelizing structure prediction on GPU SLURM-based clusters.

In the 2024 CASP16 protein structure competition, we conducted a systematic experiment involving massive sampling for all protein targets [2]. Beyond our baseline submission (group 'Brysbaert'), we contributed these extensive datasets to the CASP Phase 2 'MassiveFold scoring' challenge. This collective effort aimed to provide the community, and specifically scorers, with a diverse pool of decoys to identify the 'hidden gems' generated by MassiveFold. Our analysis highlights a sampling-scoring gap: while these high-accuracy models exist within the ensembles, they are not systematically ranked first by standard AF2 confidence scores. We show that a perfect scoring would have ranked first in CASP16 with our MassiveFold set, highlighting an urgent need for advanced scoring methods. Furthermore, our data indicates that while 'easy' targets gain little from increased sampling, 'hard' targets consistently benefit from it. A standard AlphaFold2 run can serve as a diagnostic to predict which targets require massive sampling, allowing for a significant reduction in total computing time through selective resource allocation. Finally, we present the evolution of MassiveFold (v1.6.1), which now integrates AlphaFold3, PPI discovery, ligand screening, and support for local (non-SLURM) environments.

References:

[1] Nessim Raouraoua, Claudio Mirabello, Thibaut Véry, Christophe Blanchet, Björn Wallner, Marc F. Lensink and Guillaume Brysbaert. Nature Computational Science 4, 824–828 (2024).

[2] Nessim Raouraoua, Marc F. Lensink, Guillaume Brysbaert, Proteins: Structure, Function, and Bioinformatics (2025): 1–7, (2025).

URL

https://onlinelibrary.wiley.com/doi/10.1002/prot.70040

https://github.com/GBLille/Massivefold

Cluefish: a workflow for comprehensive biological interpretation of transcriptomic data series

Oral Presentation

Ellis Franklin [1], ***Elise Billoir*** [1], ***Philippe Veber*** [2], ***Jérémie Ohanessian*** [1], ***Marie Laure Delignette-Muller*** [2], ***Sophie Prud'homme*** [1]

1. *Université de Lorraine, CNRS, LIEC, F-57000 Metz, France,* ***2.*** *Université de Lyon, CNRS, VetAgro Sup, LBBE, F-69622 Villeurbanne, France*

Abstract

Transcriptomic "data series" — such as time-course or dose-response experiments — present significant challenges in both analytical modelling and biological interpretation. While specialised pipelines exist to characterise these series, a gap remains in tools dedicated to making sense of the extensive transcript lists they generate. To address this, functional enrichment analysis is often employed to associate these lists with simplified biological pathways. A commonly used method is Over-Representation Analysis (ORA), which uses statistical tests to determine if a pathway contains a disproportionately large number of deregulated genes compared to sampling among genes uniformly. However, standard ORA often yields broad, redundant results representing only a small fraction of deregulated genes, frequently overlooking smaller, more specific pathways.

To address these limitations, we developed Cluefish (CLUstering, Enrichment, and FISHing), an open-source semi-automated R workflow for untargeted exploration of transcriptomic data series. Cluefish applies ORA to pre-clustered protein-protein interaction networks, using clusters as anchors to identify smaller and more specific pathways. The workflow incorporates two innovative steps: cluster merging to reduce functional redundancy, and "lonely gene fishing", which recovers isolated deregulated genes through shared pathway context. Together, these features enable a more complete exploration of the data by maximising the proportion of the deregulated gene list included in the interpretation.

We tested Cluefish using an in-house dose-response dataset of zebrafish embryos exposed to dibutyl phthalate (known endocrine disruptor), as well as two publicly available toxicology datasets for rat and poplar. Compared to the standard approach, Cluefish allowed for the interpretation of a larger portion of the data and revealed disrupted pathways that would otherwise be overlooked. Coupled with dose-response modelling from DRomics, Cluefish outputs enabled the formulation of hypotheses supported by multiple concordant elements, which are not only biologically coherent but also of broader scientific interest.

URL

https://doi.org/10.1093/nargab/lqaf103

Combining phenotypic similarity and network propagation to improve performance and clinical consistency of rare disease diagnosis

Oral Presentation

Maroua CHAHDIL [1], ***Carolina Fabrizzi*** [2], ***Marc Hanauer*** [2], ***Caterina Lucano*** [2], ***Ana Rath*** [2], ***David Lagorce*** [2], ***Laurent Tichit*** [3]

1. *INSERM, US14 – Orphanet, Plateforme Maladies Rares, and Aix Marseille Univ, CNRS, I2M,* ***2.*** *INSERM, US14 – Orphanet, Plateforme Maladies Rares, 96 rue Didot, 75014, Paris, France,* ***3.*** *Aix Marseille Univ, CNRS, I2M, Marseille, France*

Abstract

Achieving timely diagnosis for rare diseases (RDs) remains challenging due to phenotypic heterogeneity and incomplete clinical data. Unlike our previous phenotype–genotype-based method developed during the SOLVE-RD project (Lagorce et al. 2024), and in contrast to existing phenotype-based tools that rarely exploit Orphanet's hierarchical classification, we present a phenotype-based computational pipeline that ranks candidate ORPHAcodes (Orphanet's unique, time-stable RD identifiers) from patient phenotypes by integrating a clinical consistency metric. Following the Orphanet Nomenclature, this metric quantifies the number of shared Groups of Disorders (GDs) between two ORPHAcodes. GDs are defined as sets of ORPHAcodes sharing clinical features. More shared GDs reflect greater clinical coherence.

The pipeline (Fig. 1) combines phenotype-based similarity with Random Walk with Restart (RWR). Patient-disease similarity is computed using asymmetric aggregation of the Human Phenotype Ontology (HPO) terms, incorporating removal of subsumed terms and integration of Orphanet HPO frequency annotations. Among 24 combinations of six similarity measures and four aggregation functions, Resnik with FunSimMaxAsym performed best.

By exploring the patient node's local neighbourhood, RWR identifies indirectly related diseases through network connectivity rather than direct annotation overlap, mitigating incompletely annotated cases and improving the clinical consistency of top-ranked candidates; although some confirmed diagnoses occasionally ranked lower, their GDs appeared prominently, supporting diagnostic utility.

We benchmarked our method on 139 expert-curated SOLVE-RD cases representing 78 distinct ORPHAcodes. Compared with the SOLVE-RD baseline (RSD), our approach achieved a harmonic mean rank of 4.64 for confirmed diagnoses (versus 7.97) and retrieved the correct RD within the top 10 positions for 39% of patients (versus 29%). The damping factor ($\alpha = 0.3$) modulated the trade-off between ranking accuracy and clinical coherence.

By integrating phenotype-based similarity with RWR, our pipeline enhances clinical consistency among top-ranked candidates by considering the Orphanet nomenclature, providing a starting point to assist clinicians in developing diagnostic hypotheses.

Code: github.com/Orphanet/OrphaScape_SimWalk

URL

https://hal.science/hal-05517071v1

De novo assembly pipeline for the characterization of the Mandrillus sphinx microbiome using massive sequencing data

Oral Presentation

Raphaël Ribes [1], ***Céline Mandier*** [1], ***Elodie Flaven-Noguier*** [2], ***Fabienne Justy*** [2], ***Alice Baniel*** [2]

1. *ISDM, Univ Montpellier, CIRAD, INSERM, Montpellier, France,* ***2.*** *ISEM, Univ Montpellier, CNRS, IRD, Montpellier, France*

Abstract

The gut microbiome is fundamental to host ecology, physiology, and health, influencing key biological processes and adaptation to natural environments. Despite its importance, the microbiome of wild non-model animals remains poorly characterized. Historically, studies have relied on 16S rRNA sequencing, which only captures a fraction of microbial diversity and provides limited functional resolution. To overcome this, our study leverages high-throughput whole-metagenome sequencing to achieve a high-resolution characterization of the gut microbiome in a wild primate population.

We focus on a natural, fully habituated social group of approximately 300 mandrills (Mandrillus sphinx) in Southern Gabon's Lékédi Park, which has been continuously monitored since 2012. Analyzing 388 fecal samples collected from 86 individuals between 2021 and 2022, we generated a an extensive sequencing dataset exceeding 30 TB. Because comprehensive reference genomes for the mandrill microbiome are currently lacking, we employed a de novo assembly strategy to maximize the recovery of novel and uncharacterized native microbial taxa.

Through the reconstruction of metagenome-assembled genomes (MAGs), this project aims to produce the first reference catalog of the mandrill gut microbiome. Beyond taxonomic and functional profiling, this catalog will enable us to investigate how key socio-ecological factors such as sex, age, dominance rank, and seasonal variations shape microbiome composition and function. Furthermore, strain-level resolution and specific mutation tracking will allow us to detect clusters of microbiome transmission driven by social interactions. Ultimately, this research provides unprecedented insights into the functional potential and ecological dynamics of the gut microbiome in wild primates.

De novo genes in diatoms

Oral Presentation

Alix Boutheroue-Desmarais [1], ***Ingrid Lafontaine*** [1]

1. *Institut Biologie Physico-Chimique*

Abstract

Diatoms are the most diverse group of algae, comprising more than 100,000 species and producing about 20% of Earth's oxygen while representing nearly half of oceanic algal biomass. Their plastids originated through secondary endosymbiosis with a red alga, itself derived from a primary endosymbiotic event between a heterotrophic eukaryote and a photosynthetic bacterium. Endosymbiosis triggered extensive cellular and genomic reorganization, including gene loss from the plastid genome and transfer of many genes to the host nucleus. Diatoms have evolved innovations to accommodate this event.

To investigate the evolutionary origin of diatom genes and identify candidate *de novo* genes, we performed a large-scale phylostratigraphic analysis of 1,144,008 protein-coding genes from 58 high-quality annotated diatom genomes, complemented by 6,982 transcriptionally supported but previously unannotated genes. Overall, 95.6% of genes were grouped into 31,019 gene families.

We identified 424,781 taxonomically restricted genes (TRGs), in 3,251 families whose all members lack detectable homologues outside diatoms. Additional filters, including the absence of conserved protein domains and machine-learning classification, allowed the identification of high-confidence 120,923 *de novo* gene candidates unlikely to represent highly diverged homologues.

The ultimate evidence of de novo origination for a given gene, is to be able to identify the homologous non-coding region. This can be provided by significant similarity or by a shared microsyntenic environment. 9,236 de novo candidates share significant similarity with a non-coding region, 49 of them within a microsyntenic environment. 4,913 other de novo candidates are found within a microsyntenic region.

Functional annotation and 3D structure prediction using deep learning approaches showed that some of candidates remain uncharacterized "dark matter" genes. Among them, the one that are predicted to localize to plastids may participate in regulation, suggesting roles in diatom-specific post-endosymbiotic innovations.

Decoupling phenotypic from genetic pleiotropy during evolution on a complex genotype-phenotype-fitness map

Oral Presentation

Théotime Grohens [1], ***Marie Sémon*** [1], ***Sophie Pantalacci*** [1]

1. *LBMC CNRS UMR 5239, Ecole Normale Supérieure de Lyon*

Abstract

Pleiotropy is the fact that a gene or mutation can affect several biological functions.It can occur at several scales: ranging from expression in different tissues, to phenotypes during development, up to adult phenotypes.As a result, mutations in pleiotropic genes do not always result in pleiotropic changes at the phenotypes of interest.This is usually explained by gene modularity, in which mutations in modular enhancers enable tissue-specific variation in gene activity.But another underestimated possibility is that this decoupling also takes place at higher scales of organization.

In this work, we tackle this problem using an _*in silico*_ individual-based model of tooth development and evolution.In our model, individuals have a realistic genetic architecture based on the regulation of pleiotropic genes by both modular and pleiotropic enhancers.We compute tooth shapes based on gene expression using a classic developmental model, and fitness based on the quality of tooth occlusion (how efficiently teeth interlock with one another).Our model therefore presents a complex genotype-phenotype-fitness map, in which we can evaluate pleiotropy at several scales independently.

Using this model, we show that the _*effective pleiotropy*_ of mutations in pleiotropic genes, i.e. the phenotypic impact of mutations on both teeth, decreases during evolution, and that modular enhancers accumulate comparatively more mutations than pleiotropic ones.Moreover, we show that mutations in different genome compartment follow different evolutionary laws: while mutations in pleiotropic regions are more often deleterious than mutations in modular regions, beneficial mutations in pleiotropic regions are however on average more beneficial than mutations in modular regions.This shows how a decoupling between different levels of pleiotropy can evolve through the interaction between the modular and pleiotropic parts of the genome.

Overall, our results show how genetic architecture itself can evolve to buffer the phenotypic effect of mutations in pleiotropic genes.

Epidemics of temperate phages and what's left of them in bacterial genomes

Oral Presentation

Julien Guglielmini [1], ***Eduardo Rocha*** [2]

1. *Hub de Bioinformatique, Institut Pasteur,* ***2.*** *Microbial evolutionary genomics, Institut Pasteur*

Abstract

Bacteriophages (phages) are viruses that infect bacteria. Among them, temperate phages can establish lysogenic relationships by inserting their genetic material into the bacterial chromosome, where they persist as prophages. These integrated elements can either benefit the host through lysogenic conversion or kill it upon lytic induction. Over time, prophages can accumulate mutations that inactivate viral functions, transitioning from autonomous viruses to degraded genomic elements that can be co-opted for bacterial functions. While previous studies have documented prophage degradation and domestication, the temporal dynamics of these processes remain poorly understood.

Here, we characterize the evolutionary history of prophages in 3,783 high-quality genomes of *Escherichia coli* and *Salmonella enterica*. To detect prophages, even when severely degraded, we used geNomad with relaxed detection criteria followed by expert curation using pangenome data to remove false positives. We classified the resulting 25,950 prophages into 537 families based on gene repertoire similarity and genomic localization, and annotated the major viral functions. Finally, ancestral state reconstructions allowed us to define 1,470 clusters of orthologous prophages (COPs) as elements derived from single integration events and transmitted vertically by identifying groups of closely related prophages from the same family in the host phylogeny. This enabled analysis of the relative ages of these COPs.

Our results reveal distinct phage epidemics: waves of integrations involving similar phages occurring at comparable evolutionary times within each species. Furthermore, we show that most prophages were acquired recently, while only a few ancient elements show extensive degradation consistent with domestication. These ancient prophages lack essential viral genes yet retain signatures of purifying selection, representing transitions from parasitic elements to permanent bacterial assets that contribute to host fitness. This temporal framework provides novel insights into the balance between recent phage infections and long-term bacterial adaptation through prophage co-option.

Exploration of Multiconformers to Extract Information About Structural Deformation Undergone by a Protein Target: Illustration on the Bcl-xL Target

Oral Presentation

Marine Baillif [1], ***Eliott Tempez*** [1], ***Anne Badel*** [1], ***Leslie Regad*** [1]

1. *Université Paris Cité - BFA CNRS UMR 8251 - IsPP INSERM U1133*

Abstract

Analyzing how proteins change conformation upon ligand binding or mutations is essential for understanding molecular recognition and structure-based drug design. Traditionally, these changes are assessed by comparing apo and holo structures using RMSD. This parameter provides only a global measure of structural deviation and does not allow the localization of conformational changes. To overcome these limitations, we previously developed **SA-conf (Regad et al., 2017),** a tool that quantifies backbone conformational variability across sets of protein structures https://owncloud.rpbs.univ-paris-diderot.fr/owncloud/index.php/s/uOdItX27jUfgQM6. It is based on **HMM-SA (Camproux et al., 2004),** a structural alphabet that encodes local 3D geometry into sequences of structural letters where each encodes the geometry of consecutive four-Cα fragments.

In this study, we aimed to demonstrate the applicability of SA-conf for analyzing structural variability in Bcl-xL (B-cell lymphoma–extra-large), a key anti-cancer target. To this end, SA-conf was applied to a dataset of 130 crystallographic chains, including apo, holo, wild-type, and mutant forms. Our analysis revealed that **Bcl-xL shows** high structural plasticity. Structurally conserved positions were identified, notably in the protein fold and binding site anchor residues. While most mutations had limited local impact, some induced **long-range conformational rearrangements,** suggesting allosteric effects. We showed that ligand binding was the main driver of conformational changes, with rearrangements both near and distal to the binding site. By integrating SA-conf results with residue flexibility analysis, we established a structural mapping of the binding pocket with three important regions: (i) a **conserved anchoring core**, (ii) a**moderately plastic central region,** and (iii) a **flexible periphery** contributing to ligand specificity. These insights highlight SA-conf's ability to capture **functionally relevant backbone adaptations**. SA-conf offers a powerful framework for analyzing structural ensembles, identifying conformational signatures, and guiding **rational design of selective ligands** for dynamic protein targets. This study was previously reported in Baillif et al., Molecules, 2025

URL

https://www.mdpi.com/1420-3049/30/16/3355

Genomic analysis of the factors influencing the localization of recombination events and the segregation of genetic determinants of quality in an interspecific context in the genus Vitis

Oral Presentation

Léonie Chrétien [1], ***Camille Rustenholz*** [2], ***Guillaume ARNOLD*** [1], ***Komlan AVIA*** [1], ***Raymonde BALTENWECK*** [3], ***Patricia CLAUDEL*** [1], ***Éric DUCHÊNE*** [1], ***Philippe HUGUENEY*** [1], ***Aurélie UMAR-FARUK*** [1]

1. *INRAE,* ***2.*** *1131 SVQV, INRAE,* ***3.*** *Université de Strasbourg*

Abstract

Reducing pesticide use in viticulture is challenging, as grapevine is France's second most pesticide-consuming crop. *Vitis* vinifera, bred for wine quality, is highly susceptible to fungal diseases, while wild grapevine species with resistance genes are generally unsuitable for winemaking. Interspecific crosses have been used to introgress resistance. Since 2000, the ResDur program at INRAE Colmar has applied marker-assisted backcrossing to retain resistance while minimizing wild genome introgression, resulting in 18 registered resistant varieties. However, linkage drag remains a major challenge as undesirable traits may co-segregate with resistance genes. Meiotic recombination is key to breaking this association, yet recombination landscapes in grapevine interspecific hybrids are poorly understood.

We investigated recombination distribution in interspecific crosses, identifying hotspots and coldspots and examining their genomic features. We hypothesized that recombination is reduced in highly divergent regions and may be influenced by epigenetic features such as chromatin state.

An F1 hybrid between *V. vinifera* 'Cabernet Sauvignon' and *V. riparia* 'Gloire de Montpellier' was crossed with 'Chardonnay' to generate 190 progenies, all genotyped via Genotyping-By-Sequencing. High-density genetic maps revealed suppressed recombination in centromeric regions and substantial variation along other regions. This strategy will be extended to a more complex hybrid population with mosaic ancestry. And future work will integrate high-quality genome assemblies and berry metabolomics to link recombination with both resistance and quality traits.

These findings highlight variation in recombination landscapes, informing strategies to separate resistance loci from linked undesirable traits and guiding breeding approaches to combine disease resistance with desirable wine quality.

In situ viral regulation of bacterial successions during organic matter turnover

Oral Presentation

Domitille Jarrige [1], ***Pierre-Alain Maron*** [2], ***Eric Dugat-Bony*** [3], ***Vincent Tardy*** [2], ***Abad Chabbi*** [4], ***Olivier Rué*** [5], ***Nicolas Ginet*** [6], ***Mireille Ansaldi*** [6], ***Valentin Loux*** [7], ***Sébastien Terrat*** [8]

1. *Agroécologie, Institut Agro Dijon, INRAE, Université Bourgogne Europe,* ***2.*** *Agroécologie, Institut Agro Dijon, INRAE, Université Bourgogne Europe, F-21000 Dijon, France,* ***3.*** *Université Paris-Saclay, INRAE, AgroParisTech, UMR SayFood, Palaiseau, France,* ***4.*** *INRAE, Poitou-Charentes, URP3F, 86600, Lusignan, France ; UMR-ECOSYS Joint research unit INRAE, AgroParisTech, Université Paris-Saclay, Paris, France,* ***5.*** *Université Paris Saclay, INRAE, MaIAGE, Jouy-en-Josas, France ; INRAE, BioinfOmics, MIGALE bioinformatics facility, Université Paris-Saclay, Jouy-en-Josas, France,* ***6.*** *Laboratoire de Chimie Bactérienne,* ***7.*** *Université Paris-Saclay, INRAE, MaIAGE, 78350, Jouy-en-Josas, France; Université Paris-Saclay, INRAE, BioinfOmics, MIGALE bioinformatics facility, 78350, Jouy-en-Josas, France,* ***8.*** *Agroécologie, Institut Agro Dijon, INRAE, Université Bourgogne Europe, 17 Rue de Sully, F-21000 Dijon, France*

Abstract

Terrestrial viruses are suspected to play a significant role in top-down control of soil microbial populations, yet, direct scientific evidence of this regulation *in situ* remains scarce. To investigate the impact of soil viral communities on bacterial succession, we employed an *in situ*, temporal, global metagenomic approach, following wheat straw amendment in agricultural soils with contrasting land-use histories (20 years cropland vs. 17 years grassland). Using several bioinformatic tools to identify viral contigs (VirSorter2, geNomad, VIBRANT), approximately 300 000 putative contigs were retrieved and annotated from the 18 M contigs (>1kb) generated by the assembly. These viral sequences were highly diverse, many unknown, and displayed distinct responses both to contrasting land-use histories and to the amendment. Focusing on a subset of bacteriophage contigs with high confidence (>3kb, identification by several viral detection tools, host predictions inferred using iPHoP), we observed rapid viral dynamics mirroring those of bacterial communities. Moreover, viral/predicted host contigs ratios suggested contrasting ecological strategies; some bacteriophages seemed strongly induced, consistent with a potential kill-the-winner dynamic, whereas others appeared lysogenic, proliferating under a piggyback-the-winner model. Altogether, our findings suggest a strong role of viruses in regulating bacterial communities in soils. Furthermore, the large number of unknown viral sequences highlights the vast, yet unexplored, diversity of soil viromes awaiting discovery.

Intrinsic Conformational Dynamics of Apo HIV-2 Protease Reveal Two Dynamical Phases and Multiple Closed Flap States

Oral Presentation

Marine Baillif [1], ***Phuong Nhung Cao*** [1], ***Leslie Regad*** [2]

1. *Université Paris Cité,* ***2.*** *Université Paris Cité - BFA CNRS UMR 8251 - IsPP INSERM U1133*

Abstract

HIV-2 protease (PR2) is essential for viral maturation, making it a key therapeutic target. Experimental structures show a semi-open apo state and a closed ligand-bound state, but the intrinsic dynamics of apo PR2 remain incompletely characterized, despite their central role in enzymatic function, inhibitor recognition, and drug resistance. Here, molecular dynamics simulations combined with structural and dynamical analyses are used to investigate the conformational landscape of PR2 in the absence of ligands. Analysis of three independent trajectories reveals that PR2 undergoes a reproducible two-phase dynamical behavior. An initial dynamic phase is characterized by enhanced flexibility and extensive cooperative motions, followed by a stabilization phase associated with reduced mobility. This progression reflects PR2's shift from semi-open ensembles through transient openings toward predominantly closed conformations. To structurally dissect these conformational transitions, we developed complementary geometric metrics that separately quantify flap–flap rearrangements and residue movements relative to the active site. The detailed analysis of flap conformations demonstrates that full flap opening enables a reorientation step in which flap B moves in front of flap A, a prerequisite for subsequent closure. During the stabilization phase, PR2 predominantly samples three distinct closed conformations—extended, bent, and inward-bent. These conformational states differ in both inter-flap distance and local flap geometry. These results establish a detailed reference description of the intrinsic conformational landscape and flap dynamics of apo PR2. By defining this intrinsic reference, these findings may help elucidate how ligand binding, inhibitor specificity, and resistance-associated mutations could modulate the conformational landscape of PR2.

K-mer-based exploration of large RNA sequencing collections reveals diagnostic transcriptomic variants in acute myeloid leukemia

Oral Presentation

Chloé BESSIERE [1], ***Florence RUFFLE*** [1], ***Benoit GUIBERT*** [1], ***Ambre GALY*** [1], ***Camélia SENNAOUI*** [1], ***Anthony BOUREUX*** [1], ***Jérôme REBOUL*** [1], ***Thérèse COMMES*** [1], ***Nicolas GILBERT*** [1]

1. *Institute for Regenerative Medicine and Biotherapies (IRMB), U1183, Univ Montpellier, INSERM*

Abstract

Acute myeloid leukemia (AML) is a genetically heterogeneous clonal disorder driven by somatic alterations that disrupt proliferation, differentiation, and self-renewal of hematopoietic progenitors. Advances in next-generation sequencing technologies have considerably improved AML classification and risk stratification. In particular, RNA sequencing (RNA-seq) enables the detection of gene fusions, aberrant transcripts, and expression signatures relevant for diagnosis. However, the rapidly increasing volume of public RNA-seq datasets makes large-scale exploration computationally challenging.

To address this limitation, we develop k-mer-based strategies enabling ultra-fast interrogation of large RNA-seq collections while preserving the full information contained in sequencing reads. Our framework relies on Reindeer, an indexing method based on colored de Bruijn graphs that compress and record the abundance of all k-mers across thousands of RNA-seq libraries. This structure allows RNA sequences to be decomposed into k-mers and quantified across entire datasets within seconds. Building upon this indexing strategy, we implement Transipedia, a modular framework combining large-scale RNA-seq indexing, automated generation of event-specific k-mer signatures, and a web platform facilitating query execution and data sharing (transipedia.org).
We also develop automated strategies to construct query sequences representing transcriptomic events such as gene fusions or mutations from database annotations. Their specific k-mer signatures can be searched across indexed datasets, enabling rapid screening of candidate biomarkers without reprocessing raw sequencing data. Building on previous validation across the Cancer Cell Line Encyclopedia (CCLE, n=1019 human cancer cell lines), we extend this strategy to AML patient cohorts, including BeatAML study, and compare them with datasets from healthy donors to prioritize disease-specific alterations. This approach further enables the generation of short diagnostic sequence signatures (called *probes*), which will be aggregated into a curated and extensible repository of AML diagnostic markers for large-scale transcriptomic screening.

KGATE: A tool for graph representation learning applied to Biomedical Knowledge Graphs

Oral Presentation

Benjamin Loire [1], ***Célia Brahimi*** [2], ***Galadriel Brière*** [3], ***Anaïs Baudot*** [4]

1. *Neurology Therapeutic Area, R&D Servier Paris-Saclay Institut,* ***2.*** *Aix Marseille University, INSERM, MMG UMR 1251,* ***3.*** *Aix-Marseille University, CNRS, IBDM UMR7288, Turing Center for Living Systems (CENTURI), Marseille, France,* ***4.*** *CNRS*

Abstract

A Knowledge Graph (KG) is a data structure where entities are encoded as nodes and relations between entities as edges holding semantic meaning. A KG is often represented as a set of triplets (head, relation, tail) where head and tail are respectively the origin and destination nodes, and relation is the type of the edge connecting them.

Knowledge Graph Embedding (KGE) is a machine learning technique that learns a vectorial representation of KG nodes and edges. The most common architecture for KGE models is the autoencoder, where the encoder transforms the KG in a vectorial representation and the decoder attempts to reconstruct the original graph from the latent space.

Our library, called Knowledge Graph Autoencoder Training Environment (KGATE), is designed to address the limitations of current KGE libraries. KGATE assembles each part of an autoencoder as modular blocks that users can fine-tune. As a general-purpose KGE library, KGATE can easily compare any model in a reproducible context, focusing on fast prototyping and model benchmarking to ensure the best approach for each project.

We demonstrate the utility of KGATE in performing reproducible and explainable KGE analyses with an application on biomedical KGs. KGATE makes it possible to compare a large palette of KGE encoders and decoders and identify the best combination to predict previously unseen therapeutic targets or the repurposing of existing drugs to treat new diseases.

In addition, most biomedical KGs are static, meaning they represent a global and frozen view of a biological system, lacking temporal information such as time-dependent variations. However, biological organisms are inherently dynamic, particularly with respect to aging, and a KG intended to model an aging organism must therefore incorporate temporal properties. KGATE implements the building blocks for temporal embeddings, and handles the most common graph temporal representations to help researchers explore and develop methods.

URL

https://github.com/BAUDOTlab/KGATE

Knowledge graph mining linking endometriosis and pollutants

Oral Presentation

Meije Mathé [1], ***Guillaume Laisney*** [2], ***Olivier Filangi*** [2], ***Franck Giacomoni*** [2], ***Maxime Delmas*** [3], ***German Cano-Sancho*** [4], ***Fabien Jourdan*** [5], ***Clément Frainay*** [1]

1. *Toxalim, Université de Toulouse, INRAE, ENVT, EI-Purpan,* ***2.*** *INRAE,* ***3.*** *Idiap Research Institute,* ***4.*** *Oniris, INRAE, Laberca,* ***5.*** *MetaboHUB, National Infrastructure of Metabolomics and Fluxomics*

Abstract

Endometriosis, a chronic inflammatory disease affecting 5–15% of individuals assigned female at birth, involves the presence of endometrial-like tissue outside the uterus. While genetic and hormonal factors are established contributors, environmental exposure to persistent organic pollutants (POPs), especially organochlorine compounds (OCCs), may also play a role, though the underlying mechanisms remain unclear due to fragmented evidence in the literature.

To address this, we developed Kg4j, an open-source Java framework for constructing question-centric biomedical knowledge graphs (KGs). Kg4j leverages FORVM, a large-scale RDF graph with 82 million associations between small molecules and biomedical concepts. Using ontology-based filtering, Kg4j extracts a locally hosted Neo4j subgraph, tailored to specific research questions, providing relevant literature-supported associations.

We applied Kg4j to explore links between OCC exposure and endometriosis risk. The resulting KG (2,706 nodes, 23,243 edges) was validated against a systematic review, achieving 95.4% recall and revealing new hypothetical associations. Centrality analysis highlighted hormone-related entities and steroid metabolites, supporting the hypothesis that endocrine disruption and altered steroid metabolism contribute to pathogenesis. Entities related to reproductive and neoplastic processes also emerged as central, suggesting potential links to cancer and environmental exposures.

Exploration of chains of concepts and chemical entities within the knowledge graph, from concept of interest to central nodes, revealed multi-step connections between OCC exposure and endometriosis-related processes. These interpretable paths enable hypothesis generation about potential mechanisms through which environmental pollutants may influence endometriosis development. The framework now supports metabolomics data integration, to extract KGs driven by both data and literature. This integration allows observed metabolic disruptions to be connected to known biological pathways, improving mechanistic insight and supporting more precise hypothesis generation.

This approach demonstrates how KG-based literature mining can transform fragmented knowledge into mechanistic hypotheses, offering a scalable tool for uncovering biomarkers, pathways, and therapeutic targets in complex diseases.

URL

https://doi.org/10.64898/2026.03.02.709027

Long and small RNA annotation reveals extensive gene expression changes during chicken and pig development

Oral Presentation

Cervin Guyomar [1], ***Sarah Djebali*** [2], ***Sylvain Foissac*** [3]

1. *Sigenae, GenPhySE, Université de Toulouse, INRAE, ENVT,* ***2.*** *IRSD, UMR1220 INSERM,* ***3.*** *GenPhySE, Université de Toulouse, INRAE, ENVT, 31326, Castanet Tolosan, France*

Abstract

Functional annotation of livestock genomes is essential to understand genomic regulation and enable precision breeding. The international FAANG (Functional Annotation of ANimal Genomes) initiative was launched in 2016 to address this need. As part of this effort, the European GENE-SWitCH (GS) project aims to generate comprehensive functional genomic resources for two economically important monogastric livestock species, chicken and pig.

To characterize transcriptional dynamics during development, the GS project sampled seven tissues (liver, kidney, ileum, lung, heart, skin, and cerebellum) from four animals at three developmental stages (early embryogenesis, late embryogenesis, and newborn) in both species. Multiple sequencing assays were applied to identify transcripts, regulatory elements, and their relationships and dynamics across development.

Using state-of-the-art computational pipelines, including the TAGADA (Transcripts And Genes Assembly, Deconvolution, and Analysis) pipeline with stringent filtering and quality controls, we generated extensive annotations of long and short RNAs. Our analyses identified ~34,000 long and ~119,000 short transcripts in chicken, and ~47,000 long and ~53,000 short transcripts in pig. These annotations extend existing references such as Ensembl and RefSeq by capturing overlooked transcript classes, including numerous long non-coding RNAs, chimeric transcripts and additional transcript isoforms of known genes. Small RNA annotation using ShortStack identified known RNA classes (miRNA, tRNA, snoRNA, snRNA, rRNA, and scaRNA) as well as many transcripts of unknown biotype. Integration of RNAseq and sRNAseq data highlighted the presence of promoter associated sRNAs and sRNA expression patterns resulting from differential mRNA processing.

Transcriptome analysis revealed widespread developmental regulation. Differential expression showed that most genes—coding and non-coding—change expression between developmental stages, with 92% of chicken genes and 84% of pig genes differentially expressed in at least one tissue. Comparative analyses also identified orthologous relationships between chicken and pig genes, providing new insights into genome organization and developmental regulation in livestock species.

METAFLUX: A method for predicting metabolic fluxes by integrating proteomic data into a genome-scale constraint-based metabolic model

Oral Presentation

Maëla Sémery [1], ***Marianyela Petrizzelli*** [2], ***Sylvain Prigent*** [3], ***Mélisande Blein-Nicolas*** [4], ***Christine Dillmann*** [1]

1. *Université Paris-Saclay, INRAE, AgroParisTech, GQE – Le Moulon,* ***2.*** *Sanofi,* ***3.*** *UMR1332 BFP / MetaboHUB, PHENOME-EMPHASIS,* ***4.*** *Université Paris-Saclay, INRAE, AgroParisTech, GQE – Le Moulon ; Université Paris-Saclay, INRAE, AgroParisTech, CNRS, EMR 9005 GEvAD*

Abstract

Advances in high-throughput phenotyping now enable the characterization of many individuals at the transcriptome, proteome, and metabolome levels across different environments. However, linking genetic or molecular changes to phenotypic adaptations remains challenging. Data integration methods, particularly those using constraint-based metabolic models, can predict metabolic fluxes and provide new insights into the most active metabolic pathways under varying environmental conditions.

Several approaches predict fluxes across metabolic networks by applying constraints to solve steady-state equations. Some use omics data to identify fluxes that best fit experimental observations. Among these, Petrizzelli *et al.* proposed a method using proteomic data to constrain flux predictions, which has been tested on a simplified model of central carbon metabolism in yeast. To investigate maize adaptation to drought, we developed the METAFLUX method, derived from Petrizzelli's approach, to predict metabolic fluxes in 254 maize hybrids grown under two irrigation conditions, using available proteomic data and a curated metabolic model of a maize leaf. We tested our method on the yeast dataset, and we show its efficiency in predicting metabolic fluxes in different environmental conditions. We also confirm the high repeatability of flux predictions across 10 independent simulations. The preliminary results on a few genotypes of maize are encouraging for predicting metabolic fluxes on all samples of maize with enough precision to measure the impact of environmental conditions. This will help us to clarify maize's response to drought stress and discover which metabolic pathways contribute to drought resistance.

Multi-objective metabolic modeling of cross-feeding interactions in a microalgae-bacteria consortium under vitamin B12 stress

Oral Presentation

Marinna Gaudin [1], ***Lou Patron*** [1], ***Damien Eveillard*** [2], ***Francis Mairet*** [3], ***Enora Briand*** [1], ***Matthieu Garnier*** [1]

1. *IFREMER, PHYTOX, GENALG, 44000 Nantes, France,* ***2.*** *Nantes Université, École Centrale Nantes, CNRS, LS2N, UMR 6004, 44000 Nantes, France,* ***3.*** *IFREMER, PHYTOX, PHYSALG, 44000 Nantes, France*

Abstract

Marine microalgae, key components of phytoplankton, coexist with their microbiome within the phycosphere, a microenvironment enriched in nutrients and prokaryotes in which the web of interactions is particularly intense. Together, they form a holobiont (or host-microbiome) system, acting as an integrated functional unit. *Prymnesium parvum* is a toxic haptophyte microalga responsible for harmful algal blooms (HABs) which can lead to significant ecological consequences. However, its ecology remains poorly understood, in particular the mechanisms underlying its rapid proliferation and toxicity. Notably, *P. parvum* is auxotroph for vitamin B12 (or cobalamin), an essential cofactor involved in several metabolic processes, including the cobalamin-dependent methionine synthase (metH). In environments with limited amounts of cobalamin, the alga likely relies on its microbiome bacterial partners to obtain this vitamin or metabolic alternatives to sustain growth. Indeed, metabolite exchanges, also known as cross-feeding exchanges, were shown to play a key role in shaping microbial communities dynamics and may represent relevant mechanisms for *P. parvum* to cope with cobalamin limitation. To investigate the feasibility of such interactions, genome-scale metabolic models (GSMs) were reconstructed for *P. parvum* and 15 bacterial species forming a synthetic consortium previously shown in co-culture experiments to support algal growth under cobalamin depletion. Multi-objective optimization was then used to simulate the simultaneous growth of interacting species and explore metabolic trade-offs from shared and competing resource use within nutrient-limited environments. Preliminary predictions indicate that some bacterial partners may supply cobalamin directly, while others may provide methionine, potentially bypassing the need for the cobalamin-dependent methionine synthase and alleviating B12 limitation. Overall, this multi-objective modeling framework aims to shed light on potential cooperative cross-feeding interactions and competition dynamics within holobiont systems and generates testable hypotheses to guide future experimental validation and improve our understanding of microalgal–microbiome metabolic dynamics.

NanoVar: a comprehensive workflow for structural variant detection to uncover the genome's hidden patterns.

Oral Presentation

Asmaa Samy Samy [1], Cheng Yong Tham [2], Matthew Dyer [1], Touati Benoukraf [1]

1. *Division of BioMedical Sciences, Faculty of Medicine, Memorial University of Newfoundland.,* ***2.*** *Cancer Science Institute of Singapore, National University of Singapore.*

Abstract

Structural variants (SVs), including insertions, deletions, duplications, inversions, and translocations, represent a major source of genomic variation and play a critical role in evolution, genetic diversity, and disease susceptibility. Despite their biological importance, accurate detection of SVs has historically been challenging due to their size, complexity, and frequent occurrence within repetitive regions of the genome. The emergence of long-read sequencing technologies such as Oxford Nanopore Technologies and PacBio has significantly improved the ability to resolve these complex genomic rearrangements, yet robust and accessible analytical frameworks remain essential to fully exploit these datasets.

Here we present NanoVar, a comprehensive and user-friendly workflow for structural variant detection from long-read sequencing data. NanoVar integrates efficient read alignment, candidate SV detection, neural-network–based filtering, and automated annotation within a streamlined computational pipeline. Designed to operate with modest computational resources, NanoVar enables reliable identification of structural variants from whole-genome long-read sequencing datasets within approximately 2–5 hours after read mapping. The workflow supports multiple study designs, including single-sample analyses, cohort-based studies, and genome instability investigations, and provides downstream tools for SV filtering, visualization, and annotation.

NanoVar has been successfully applied in a wide range of genomic studies, including investigations of rare genetic disorders, cancer genomics, population genomics, and genome analyses in non-model organisms. Its open-source implementation and modular architecture facilitate integration with existing genomic analysis pipelines and make the protocol accessible to researchers with limited command-line experience. The recently published Nature Protocols article provides detailed step-by-step instructions covering installation, data preprocessing, SV detection, result interpretation, and integration with complementary structural variant tools.

URL

https://www.nature.com/articles/s41596-025-01270-5

Pangenome Graph Node-Phenotype Association shows GWAS-like quality results but using only few individuals

Oral Presentation

Camille Carrette [1], ***François Sabot*** [2], ***Cédric Muller*** [3]

1. *Université de Montpellier,* ***2.*** *IRD montpellier,* ***3.*** *Syngenta Toulouse*

Abstract

Purpose: We introduce GraNPA, Graph Node-Phenotype Association, a method performing a GWAS-like analysis on a pangenome graph built with few individual sequences without the need for additional population materials and kinship information. This method reduces the number of individuals required for association studies and prevents reference bias from variant calling in these types of analysis.

Methods: A pangenome graph represents the multiple alignment of a set of complete genomes and contains all variations, from single SNPs to large structural variations, as nodes in the graph. By integrating phenotype information into nodes for each individual that crosses them, we can assign a phenotype score to each node by allocating a phenotype value to individuals based on whether or not they carry the trait, and computing the score depending on the individuals that cross nodes. This allows us to identify phenotype spots directly in the graph. These spots represent significant shifts in phenotypic score distribution, and a statistical test is used to highlight these shifts and reassign them at the chromosome level using the graph node topology.

Results: This method was tested using two publicly available datasets: the Sub1 loci and the submergence trait in Oryza sativa, with 13 individuals, and the insertion responsible for white-headed cattle, with 24 individuals. These datasets were chosen to test the method on two different species with different numbers of individuals.

Conclusion: GraNPA is able to perform the analysis with only a few dozen complete genomes in a pangenome graph and identify the nodes associated to the traits. GraNPA is still in development, for now, this method works only on qualitative phenotypes and with GFA graph containing Walks.

URL

https://forge.ird.fr/diade/graphgwas/granpa

Phage evolutionary relationships emerge from protein language model-based proteome representation

Oral Presentation

Swapnesh Panigrahi [1], ***Mireille Ansaldi*** [1], ***Nicolas Ginet*** [1]

1. *Laboratoire de Chimie Bactérienne*

Abstract

Bacteriophage taxonomy remains a challenging task due to the propensity of viruses for recombination and the lack of universal gene markers. Additionally, the rapid expansion of metagenomic datasets from diverse niches demands reliable and scalable approaches for classifying both known and novel phages. In this study, we introduce **Hierarchical Viruses (HieVi)**, a scalable framework that leverages protein language model (pLM) embeddings to compare and classify phages at the proteome level.

HieVi employs the ESM-2 model to generate vectorial representations for every protein in a phage genome. By treating phages as a "bag of proteins", we create a mean phage representation (MPR) in a 2,560-dimensional space. Using a curated dataset of 24,362 phages (INPHARED), we demonstrate that MPRs cluster effectively at the genus level. Hierarchical density-based clustering of these MPRs further reveals a multi-scale organization of phages that aligns remarkably well with ICTV taxonomic rankings, achieving an Adjusted Mutual Information score greater than 0.9 for families such as *Herelleviridae* at both genus and subfamily levels. Notably, HieVi demonstrates that highly mosaic lambdoid phages are clustered within a single, distinct branch and supports the recent ICTV promotion of *Autographiviridae* to the rank of order (*Autographivirales*). These results demonstrate that pLM-based representations can capture evolutionary relationships without relying on multiple sequence alignments.

Unlike traditional tools, HieVi is highly scalable because its vector-based formulation can leverage vector databases for rapid lookups. This enables contextual placement of new query genomes without reprocessing the entire database, facilitating the high-throughput analysis of large metagenomic datasets. The code and package are publicly available on Zenodo and GitHub, ensuring the tool is accessible and reusable. This framework shows that phage representation using protein language models contains phylogenetic information by encoding shared genes between phages, thus paving the way for advanced synteny-aware genome language models.

URL

[article] https://doi.org/10.1093/nargab/lqaf134

[App] https://huggingface.co/spaces/pswap/hievi
[Visualization] https://pswapnesh.github.io/HieVi/imgvr_galaxy.html

phyloDS: An RNA-Seq Data-Driven Method for Differential Splicing Analysis Across Species

Oral Presentation

Arnaud Liehrmann [1], ***Louis Carrel Billiard*** [1], ***Mélina Gallopin*** [2], ***Paul Bastide*** [3], ***Hugues Richard*** [4], ***Élodie Laine*** [5]

1. *Department of Computational, Quantitative, and Synthetic Biology (CQSB), UMR 7238, IBPS, Sorbonne Université, CNRS,* ***2.*** *Institute for Integrative Biology of the Cell (I2BC), Université Paris-Saclay, CEA, CNRS,* ***3.*** *Université Paris Cité, CNRS, MAP5,* ***4.*** *Genome Competence Center (MF1), Robert Koch Institute, Nordufer 20, 13353,* ***5.*** *Department of Computational, Quantitative, and Synthetic Biology (CQSB), UMR 7238, IBPS, Sorbonne Université;Institut universitaire de France (IUF)*

Abstract

Identifying the genetic mechanisms underlying phenotypic evolution is a central goal of evolutionary biology. While comparative transcriptomics is widely used to characterize lineage-specific shifts in gene expression, post-transcriptional regulation, such as alternative splicing, remains largely unexplored across species. Current methods proceed in an ad hoc manner: they are either reference-based, or they fail to adequately account for phylogenetic dependence between species. As a result, two major challenges remain unresolved: defining un-biased orthologous splice junctions and controlling false positives arising from the phylogeny. We present phyloDS, an RNA-seq driven framework for reference-free and robust differential splicing analysis across species. phyloDS combines multiple-sequence-alignment-based orthology mapping of splice junctions with phylogenetically informed differential testing via phyloDE, a regularized Ornstein-Uhlenbeck linear model approach. The method quantifies lineage-associated shifts in junction usage between predefined species groups, supports empirical null simulations to assess type I error rate control, and exports integrative visualization files for sequence- and annotation-aware interpretation. Applied to brain transcriptomes from primates and heart transcriptomes from rodents, phyloDS detected 234 lineage-associated splicing shifts across 158 genes in hominoids and 502 such shifts across 322 genes in mole-rats. Among these significant junctions 86 and 141 were unannotated in at least one species in the primate and rodent datasets, respectively, and involved exon skipping, alternative splice site usage, and alternative transcript start and end events. Moreover, differentially spliced junctions identified by phyloDS from RNA-Seq data were consistent with junctions identified from genomic sequences alone by OpenSpliceAI, supporting a cis-regulatory basis for most of the detected events. phyloDS will benefit biologists working on comparative transcriptomics by providing a robust, versatile and practical framework to detect, validate, and interpret evolutionary changes in alternative splicing from RNA-seq data.

PLM-View : Protein Language Models for fast, accurate, interpretable functional classification

Oral Presentation

Vinh-Son Pho [1], ***Alessandra Carbone*** [1]

1. *Sorbonne Université, CNRS, IBPS, UMR 7238, Department of Computational, Quantitative and Synthetic Biology (CQSB)*

Abstract

Functional classification of protein sequences remains a major bottleneck in biology. Although sequence databases have grown exponentially, most proteins remain functionally uncharacterized, particularly at the reaction level. Recent advances in Protein Language Models (PLMs), trained on massive protein sequence datasets, have shown that these models capture meaningful statistical patterns in the "protein language." Models such as ESMFold demonstrate that structural information can be inferred directly from sequence, raising the question of whether PLMs can also enable large-scale functional annotation.

Here we introduce PLM-View, an unsupervised method that leverages PLM embeddings to construct hierarchical functional classifications directly from protein sequences. PLM-View requires no additional model training and rapidly organizes large protein families into functional dendrograms. Importantly, the method provides residue-level explanations that highlight sequence positions driving functional separation.

The PLM-View pipeline is designed for computational efficiency, with most steps parallelized on GPUs. This enables rapid analysis of large protein families; in practice, functional classifications of thousands of sequences can be constructed within minutes to hours depending on dataset size. Such efficiency makes the approach suitable for large-scale exploration of modern protein sequence databases.

We applied the method to visual opsins, where PLM-View correctly identifies known functional subfamilies and resolves finer distinctions by separating sequences according to their peak absorption wavelength (lambda_max). Residue-level explanations align with experimentally characterized spectral tuning sites that determine wavelength sensitivity.

These results demonstrate that PLM representations encode rich functional information that can be extracted without additional training. PLM-View therefore provides a scalable and interpretable framework for functional annotation across the rapidly expanding universe of protein sequences.

Revisiting SIF abstraction rules with SPARQL for querying BioPAX

Oral Presentation

Cécile Beust [1], ***Olivier Dameron*** [1], ***Nathalie Théret*** [2], ***Emmanuelle Becker*** [1]

1. *Univ Rennes, Inria, CNRS, IRISA - UMR 6074, Rennes, F-35000, France,* ***2.*** *Univ Rennes, Inria, CNRS, IRISA - UMR 6074, Rennes, F-35000, France ; Univ Rennes, Inserm, EHESP, Irset, UMR S1085,Rennes, F-35000, France*

Abstract

BioPAX (Biological PAthway eXchange) is a Semantic Web-based standard for representing biological pathways, offering fine-grained descriptions of molecular interactions, metabolic processes, and signaling networks. However, BioPAX' complexity poses challenges for downstream analyses and reasoning tasks, requiring abstraction methods to simplify representation and analysis. The Simple Interaction Format (SIF) is a widely used format addressing this issue by reducing BioPAX's mechanistic interactions into binary relationships between proteins and small molecules. Currently, Paxtools and ChiBE are the state-of-the-art tools for abstracting BioPAX to SIF, relying on 14 abstraction rules implemented as Java graph patterns. However, SIF rules are ambiguously documented, leading to inconsistencies and potential misinterpretations.

Our first contribution is a systematic analysis of SIF rules ambiguity, revealing three meanings of SIF rules descriptions: (1) diagrams and textual descriptions from PathwayCommons, (2) Paxtools pattern function descriptions (comments in the code), and (3) raw Paxtools pattern Java code. We identified that each subsequent meaning introduces additional constraints not specified in previous meanings, creating a gap between SIF rules documentation and implementation in Paxtools.

Our second contribution is the development of a transparent and reproducible BioPAX-to-SIF abstraction method introducing 14 SPARQL queries formalizing each SIF rule, and clearly documented through diagrams. Using our SPARQL queries, we quantified the impact of SIF rules ambiguity on a pathway example.

Our third contribution is the production of large scale SIF abstractions of two pathway databases, PathBank and Reactome, demonstrating scalability and efficiency of our queries.

Overall, our SPARQL-based approach provides a transparent, scalable and FAIR-compliant method for BioPAX-to-SIF abstraction. While Paxtools and ChiBE remain valuable for BioPAX manipulation, our queries offer a modular, well-documented alternative for users seeking clarity in SIF abstraction rules. The scalability and efficiency of the SPARQL queries demonstrate their potential for large-scale pathway analysis. SPARQL queries are publicly available on GitHub: https://github.com/CecileBeust/BioPAX-To-SIF-SPARQL.git

URL

https://zenodo.org/records/18980848

SIDURI: an integrated data and analysis portal supporting data-driven innovation in food fermentation

Oral Presentation

Emilie Fernandez [1], ***Agnès Barnabé*** [1], ***Erwan Le Floch*** [1], ***Thomas Lacroix*** [2], ***Jonathan Mineau*** [3], ***Sophie Schbath*** [4], ***Valentin Loux*** [4]

1. *Université Paris-Saclay, INRAE, MaIAGE, 78350, Jouy-en-Josas, France; INRAE, BioinfOmics, MIGALE bioinformatics facility, 78350, Jouy-en-Josas, France; Ferments du Futur (US INRAE 1503), 91400, Orsay, France,* ***2.*** *Université Paris-Saclay, INRAE, MaIAGE, 78350, Jouy-en-Josas, France; Université Paris-Saclay, INRAE,* ***3.*** *INRAE, DipSO (Direction pour la Science Ouverte / Directorate for Open Science),* ***4.*** *Université Paris-Saclay, INRAE, MaIAGE, 78350, Jouy-en-Josas, France; Université Paris-Saclay, INRAE, BioinfOmics, MIGALE bioinformatics facility, 78350, Jouy-en-Josas, France*

Abstract

The French Grand Challenge « Ferments du Futur » (FdF) is a public-private partnership ongoing since 2022. It aims to shift from empirical to data and knowledge driven design of fermented foods. Projects funded within FdF generate heterogeneous datasets spanning microbial ecology, sensory and biochemical characterization, bioprocess engineering and host–microorganism interactions. Managing and analysing these data across academic and industrial partners requires robust digital infrastructures.

To address this challenge, the ontology-driven information system OpenSILEX [1] was selected to promote good digital practices in line with the FAIR principles. An FdF instance, called Siduri, has been developed in close collaboration with the OpenSILEX core team. Several features implemented first for Siduri are contributed back to the shared open-source codebase. Current developments focus on the management of genetic resources such as microbial strains, which play a central role in fermentation research. Developed and hosted by the Migale Bioinformatics Core Facility, Siduri centralizes FdF project results and integrates relevant public resources. However organizational solutions are needed and a comprehensive data stewardship agenda has therefore been designed to support the FdF projects along the data life cycle.

The Siduri portal also provides bioinformatics analyses with a catalogue of tools and workflows built from metadata extracted from bio.tools registry, supported by ELIXIR Europe, and the EDAM ontology. This catalogue is shared through a dedicated Galaxy instance that ensures execution and traceability of analyses. Within the Siduri portal, a ShinyProxy server deploys an interactive application that enables users to transfer datasets and their provenance between Siduri and Galaxy via their APIs.

Siduri illustrates how existing open-source software and resources can be extended and integrated to support collaborative, data-driven research in food fermentation.

1. Neveu P, Tireau A, et al. Dealing with multi-source and multi-scale information in plant phenomics: the ontology-driven Phenotyping Hybrid Information System. New Phytol. 2019;221(1):588–601.

URL

https://siduri.migale.inrae.fr/

Sister-chromatid analysis to study strand-specific DNA methylation maintenance during DNA replication

Oral Presentation

Manon Coulée [1], ***Nora Fajri*** [1], ***Antoine Pigeon*** [1], ***Nataliya Petryk*** [1]

1. *Université Paris-Saclay, CNRS, Gustave Roussy, Genome Integrity and Cancers*

Abstract

Genome replication is a finely organized process essential for cell division. It is an asymmetric process with the leading strand synthesized continuously and the lagging strand synthesized discontinuously as Okazaki fragments. To study this process, we developed new sequencing methods, SCAR-seq and OK-seq. OK-seq, which sequences Okazaki fragments, determines the direction of replication forks and location, and efficiency of replication origins across the genome. SCAR-seq reveals the occupancy of replication-associated proteins on the newly-replicated strands in a strand-specific manner.

Here, we focus on the regulation of replication-coupled DNA methylation maintenance, which consists of the remethylation of newly-synthesised strands symmetrically to parental ones. This process is essential for maintaining epigenetic regulation across cellular divisions.

We investigated the strand-specific occupancy of key DNA methylation maintenance factors on newly-synthesized strands in mouse embryonic stem cells to understand the regulation of DNA methylation maintenance during genome replication. We demonstrate that these factors are associated with the lagging strand. Since DNA methylation is essential for silencing transposable elements (TEs), we investigated the enrichment of these factors at DNA replication forks within TE regions and found that DNA methylation maintenance preferentially associates with TEs during lagging strand synthesis. These proteins appear to be particularly enriched at several types of transposable elements, such as LINEs and SINEs. We found that LINEs and SINEs are predominantly oriented head-on relative to DNA replication forks, indicating that these elements have more frequently transposed into the lagging strand during DNA replication. Moreover, the lagging strand favors the retention of intact LINEs and promotes centromeric satellite expansion in the genome over evolutionary time, suggesting that TEs have developed a strategy to hijack the lagging strand replication.

Together, our findings uncover a replication-coupled mechanism that links lagging-strand synthesis to TE silencing, with fundamental implications for genome integrity and evolution.

Spatial transcriptomic analysis reveals region-specific glial activation during epileptogenesis

Oral Presentation

Adrien Dufour [1], ***Christophe Le Priol*** [1], ***Baptiste Porte*** [1], ***Ronan Jouanard*** [1], ***Julien Maurizio*** [2], ***Anne-Elodie Receveur*** [1], ***Stéphane Auvin*** [1], ***Juliette Van Steenwinckel*** [1], ***Pierre Gressens*** [1], ***Andrée Delahaye-Duriez*** [1]

1. *INSERM,* ***2.*** *INOVARION*

Abstract

Temporal lobe epilepsy (TLE) is a common neurological disorder that often develops after an initial brain insult such as status epilepticus (SE). The transition from a healthy brain to a chronically epileptic state, known as epileptogenesis, involves complex molecular and cellular remodeling across multiple brain regions. Although transcriptomic studies have provided important insights into this process, most approaches rely on bulk or single-cell sequencing methods that lack spatial context. Here, we use spatial transcriptomics combined with integrative bioinformatic analyses to characterize the spatiotemporal transcriptional landscape of epileptogenesis.

Spatial transcriptomic profiles were generated using the 10x Genomics Visium platform from coronal brain sections of a lithium–pilocarpine rat model of SE at four time points (5, 10, 20, and 40 days post-induction), covering both latent and chronic phases. The dataset includes 16 samples and thousands of spatial capture spots per section, each representing the transcriptome of local cellular neighborhoods. Sequencing data were processed using Space Ranger and analyzed in R using Seurat, including SCTransform normalization, multi-slice integration, unsupervised spatial clustering, differential expression analysis, and pathway enrichment. Spatial cell composition was inferred using reference-based deconvolution with CARD and a mouse brain single-cell atlas.

Unsupervised clustering reconstructed major anatomical structures, including cortex, hippocampus, thalamus, white matter tracts, and caudate–putamen, demonstrating that transcriptomic signatures alone can recover brain architecture. Differential expression analysis identified more than 8,000 genes significantly dysregulated following SE, with enrichment in pathways related to immune activation, synaptic remodeling, and neuronal plasticity.

Spatial pathway analysis revealed pronounced reactive astrogliosis and microglial activation extending beyond the hippocampus into white matter tracts and thalamic nuclei during the latent phase of epileptogenesis. These results highlight the value of spatial transcriptomics and integrative bioinformatics workflows for uncovering region-specific molecular programs underlying neurological disease progression.

URL

https://doi.org/10.1186/s40478-026-02224-y

Spatiotemporal regulation of cell cycle states within the complex tumor microenvironment

Oral Presentation

Gianni Zanardelli [1], ***Olivier Tassy*** [1], ***Maulik Nariya*** [1], ***Nacho Molina*** [2]

1. *IGBMC,* ***2.*** *IE university*

Abstract

Dysregulation of the cell cycle is a hallmark of cancer, driving uncontrolled proliferation and contributing to tumor progression, metastasis, and therapy resistance. In addition to active cycling, many tumor cells can enter a quiescent (G0) state, halting proliferation without undergoing terminal differentiation. Quiescence has emerged as a key mechanism of therapeutic failure, allowing subpopulations of cancer cells to survive treatment and later initiate relapse. Furthermore, the tumor microenvironment (TME) plays a pivotal role in shaping cancer cell behavior, including their proliferative potential. Composed of a heterogeneous mix of stromal, immune, and malignant cells, the TME provides spatially organized signals that influence cell fate decisions. While immune and metabolic dimensions of the TME have been extensively studied, its role in regulating cell cycle states remains poorly understood. Particularly, the spatial distribution of proliferating versus quiescent tumor cell states, and how this organization contributes to intratumoral heterogeneity and therapeutic resistance, has yet to be systematically investigated.

In this work, we use publicly available spatial transcriptomics datasets to characterize the effect of the TME on cell cycle states in melanoma, breast, and lung tumors. Based on transcriptional profiles, we assigned discrete cell cycle phases to the cells within the TME. Interestingly we identified proliferating and quiescent subpopulations of cancer cells in the tumor. We further examined the spatial localization of these subpopulations and uncovered clear spatial patterning based on proliferating and quiescent cellular states. To refine this understanding, we develop CyclEM, a computational tool based on the Expectation-Maximization algorithm, which performs trajectory analysis, assigns a continuous cell cycle phase to proliferating cells, and reveals gene expression dynamics of key cell cycle regulators within the TME. Our approach enables a spatially resolved, high-resolution view of tumor cell cycle dynamics, uncovering heterogeneity in proliferation potential that may underlie differential responses to therapy and disease progression.

URL

https://www.biorxiv.org/content/10.64898/2025.11.28.691200v1

Supporting Workflow Reproducibility by Linking Bioinformatics Tools across Papers and Executable Code

Oral Presentation

Clémence Sebe [1], ***Olivier Ferret*** [2], ***Aurélie Névéol*** [1], ***Sarah Cohen-Boulakia*** [1]

1. *Université Paris-Saclay, CNRS, Laboratoire Interdisciplinaire des Sciences du Numérique,* ***2.*** *Université Paris-Saclay, CEA, List, F-91120*

Abstract

Motivation: The rapid growth of biological data has intensified the need for transparent, reproducible, and well-documented computational workflows. The ability to clearly connect the steps of a workflow in the code with their description in a paper would improve workflow understanding, support reproducibility, and facilitate reuse. This task requires the linking of Bioinformatics tools in workflow code with their mentions in a published workflow description.

Results: We present CoPaLink, an automated approach that integrates three components: Named Entity Recognition (NER) for identifying tool mentions in scientific text, NER for tool mentions in workflow code, and entity linking grounded on Bioinformatics knowledge bases. We propose approaches for all three steps achieving a high individual F1-measure (84 - 89) and a joint accuracy of 66 when evaluated on Nextflow workflows using Bioconda and Bioweb Knowledge bases. CoPaLink leverages corpora of scientific articles and workflow executable code with curated tool annotations to bridge the gap between narrative descriptions and workflow implementations.

Availability: The code is available at https://gitlab.liris.cnrs.fr/sharefair/copalink-experiments and https://gitlab.liris.cnrs.fr/sharefair/copalink. The corpora are also available at https://doi.org/10.5281/zenodo.18526700, https://doi.org/10.5281/zenodo.18526760 and https://doi.org/10.5281/zenodo.18543814.

URL

Link to preprint : https://arxiv.org/abs/2603.08195

Software CoPaLink : https://gitlab.liris.cnrs.fr/sharefair/copalink

Unifying genetic differentiation statistics: mathematical constraints and application to tumour evolution

Oral Presentation

Yuliya Lim [1], ***Noah Rosenberg*** [2], ***Nicolas Alcala*** [1]

1. *Computational Cancer Genomics Team, Genomic Epidemiology Branch, International Agency for Research on Cancer/World Health Organization,* ***2.*** *Department of Biology, Stanford University*

Abstract

Measures of genetic differentiation, such as F_{ST}, G'_{ST} and Jost's D are widely used to quantify evolutionary divergence and the partitioning of genetic diversity among populations. The concept of biodiversity is increasingly applied in biomedicine to describe cellular diversity. In particular, intra-tumoral heterogeneity, arising from genetically distinct tumour clones, is recognized as a key predictor of treatment response. However, the mathematical properties and comparability of the diversity statistics remain debated, limiting their clinical translation.

F_{ST}, G'_{ST} and D range from 0 to 1, with values near 0 indicating similar genetic composition and values near 1 indicating strong differentiation. However, several studies have shown that these statistics are constrained by underlying mathematical properties. F_{ST} is sensitive to within-subpopulation diversity and can remain near 0, even when populations differ substantially in genetic composition. G'_{ST} and D were designed to overcome this limitation, yet all three statistics remain constrained by total population diversity, determined by the frequency of the most frequent allele M.

In this study, we derive the allelic configurations that maximize F_{ST}, G'_{ST} and D for a fixed M and provide analytical expressions for their upper bounds. We show that for most values of M, all three statistics are constrained below 1 and reach their maximum under the same allelic configurations.

We illustrate these results using genomic data from seven patient-derived tumour organoid models sampled at consecutive time points. Standardizing each statistic by its theoretical upper bound reveals consistent differentiation levels across experiments, despite substantial differences in raw values. The statistics also reveal tumour evolution patterns: higher differentiation is observed in samples carrying damaging mutations in tumour suppressor genes, consistent with selective sweeps that substantially alter tumour genetic composition.

These findings provide a unified framework for interpreting differentiation statistics and demonstrate their utility for quantifying tumour evolutionary dynamics in cancer.

URL

https://github.com/nalcala/PopGenBounds

Updating and using a Hidden Markov Models-based algorithm to detect Anti-Microbial Resistance sequences in French soils metagenomes

Oral Presentation

Zéphyrin Enaux [1], ***Domitille Jarrige*** [2], ***Olivier Rué*** [1], ***Solène Perrin*** [2], ***Maxime Courcelle*** [1], ***Corinne Cruaud*** [3], ***Patrick Wincker*** [4], ***Claudy Jolivet*** [5], ***Samuel Dequiedt*** [2], ***Antonio Bispo*** [5], ***Lionel Ranjard*** [2], ***Valentin Loux*** [1], ***Sébastien Terrat*** [2]

1. *Université Paris-Saclay, INRAE, MaIAGE, 78350, Jouy-en-Josas, France; Université Paris-Saclay, INRAE, BioinfOmics, MIGALE bioinformatics facility, 78350, Jouy-en-Josas, France,* ***2.*** *Agroécologie, Institut Agro Dijon, INRAE, Université Bourgogne Europe, 17 Rue de Sully, F-21000 Dijon, France,* ***3.*** *Genoscope, Institut François Jacob, CEA, CNRS, Univ Evry, Université Paris-Saclay, Evry, France,* ***4.*** *Génomique Métabolique, Genoscope, Institut de Biologie François-Jacob, CEA - CNRS - Univ. Evry / Université Paris Saclay, 91000 Evry, France.,* ***5.*** *INRAE, Info&Sols, 45000 Orléans, France*

Abstract

Anti-Microbial Resistance (AMR) is identified as one of the greatest worldwide health threats for the next decades. Experts predict that AMR will cause 10 million deaths per year in 2050. Consequently, it is crucial to improve global knowledge of AMR diversity and emergence. Soils microorganisms have historically been the source of many antibiotics, and soils are an incredible reservoir of microbial biodiversity: one gram of soil can contain about a million of different species. Moreover, the excessive use of antibiotics in agriculture has led to the spread of AMR in soils. In this context, global metagenomics would allow investigation of AMR diversity, and of its ecological and evolutionary dynamics.

To this end, we assembled, processed and analysed 200 soils metagenomes, from the second sampling campaign of the French Soil Quality Monitoring Network (RMQS2), the most extensive and without *a priori* soil sampling survey available to date. We used Meta-MARC, a predictive tool based on MEGARes, an acyclic hand-curated AMR sequences database, to identify potential AMR sequences in obtained metagenome contigs. Meta-MARC uses Hidden Markov Models (HMMs), providing higher sensitivity than alignment-based algorithms to detect more divergent AMR sequences. However, Meta-MARC available version models are based on MEGARes v1.0, which is now outdated. To overcome this problem and also access information about multiresistance and metal resistance genes, we performed an update of Meta-MARC with MEGARes latest version.

Our new Meta-MARC version was built with 102 733 sequences belonging to 49 different resistance classes. First analyses of the 200 metagenomes with the updated Meta-MARC indicate links between soil use and abiotic soil parameters, and AMR classes abundance. We are implementing the Meta-MARC update process as a reproductible workflow, which will be publicly available and could be used to upgrade it with any future version of MEGARes.

URL

DOI: https://doi.org/10.57745/8Q6R92

Visualization-driven pipeline for drug design through generative AI

Oral Presentation

Lucas ROUAUD [1], ***Etienne REBOUL*** [1], ***Isleme KHALFAOUI*** [1], ***Malek MELLITI*** [1], ***Antoine TALY*** [1], ***Marc BAADEN*** [1]

1. *IBPC, LBT, UMR 8266, CNRS, UPCité*

Abstract

Designing and optimizing ligands for a known molecular target remains challenging, in part because generative methods often produce candidates that are hard to interpret. We present an interpretable pipeline which combine a synthesis-aware generative AI with two visualization tools to design new ligands. Starting from a receptor structure with a known ligand, the pipeline extracts ligand pharmacophore features and computes interaction fields once for the receptor binding pocket using smiffer. These fields show possible interaction inside a given pocket like hydrophobic or pi stacking. A Monte-Carlo tree-search generator then assembles molecular building blocks via known reactions encoded as SMARTS/SMILES, ensuring that molecules are synthetically feasible. Candidate molecules are evaluated by docking and by pairwise interaction analysis performed with strange, which computed ligand–receptor pharmacophore features, then possible interaction between these features. The pipeline produces docking scores, interaction fields, and direct interaction to facilitate interpretation. All components (smiffer, strange, and the pipeline) are open source and available online (GitLab, PyPI). The workflow is operational and currently being evaluated on ongoing benchmark, which aims to quantify the pipeline robustness.

URL

- https://gitlab.galaxy.ibpc.fr/rouaud/smiffer
- https://gitlab.galaxy.ibpc.fr/rouaud/strange

Oral Full Proceedings

Benchmark Bias and Conformational Dynamics in Allosteric Site Prediction

Victor Pryakhin[1], Malika Smail-Tabbone[1], Yasaman Karami[2,*]

[1] Université de Lorraine, CNRS, Inria, LORIA, F-54000 Nancy, France
[2] Université de Lorraine, Inria, F-54000 Nancy, France
[*] corresponding author: yasaman.karami@inria.fr

Abstract

Allosteric site prediction plays a critical role in modern drug discovery, offering opportunities to target regulatory regions with high specificity. However, most existing computational approaches rely on static protein structures and pocket detection tools such as *fpocket*, thereby overlooking conformational dynamics essential for allosteric regulation. Here, we present AlloDyn, a framework that integrates static pocket descriptors with dynamic features derived from both all-atom molecular dynamics (MD) simulations of apo-state proteins and AlphaFlow-generated conformational ensembles. By capturing structural flexibility, solvent accessibility, and residue–residue communication patterns at the pocket level, our approach enables a dynamic-aware representation of candidate allosteric sites. Importantly, we identify a systematic bias in current benchmarking practices, showing that applying *fpocket* to holo structures without removing bound allosteric modulators introduces data leakage and leads to artificially inflated performance estimates. When evaluated on properly preprocessed datasets, dynamic feature augmentation significantly improves prediction performance over static baselines. Furthermore, we demonstrate that AlphaFlow-generated ensembles achieve performance comparable to MD-derived features at a fraction of the computational cost, providing a scalable alternative for conformational sampling. Benchmarking on the D24 dataset shows that AlloDyn achieves the best balance between precision and recall, yielding the highest F1 score and MCC among evaluated methods. We show that current benchmarks overestimate performance due to data leakage, and that incorporating dynamics is key to accurate and scalable allosteric site prediction.

Introduction

Allostery is a fundamental regulatory mechanism in which the binding of a molecule (*effector* or *modulator*) at one site of a protein (the *allosteric site*) induces functional changes at a distant site, often the *active site*. This long-range intramolecular communication enables proteins to modulate their activity in response to cellular signals. Allosteric regulation is widespread across biological systems and plays a central role in controlling enzymatic catalysis, metabolic pathways, and signal transduction [1]. Importantly, effectors can act as activators or inhibitors, finely tuning protein function without directly interacting with the substrate-binding site [2].

Allosteric sites often lie outside conserved active regions, making them attractive yet challenging targets for drug design. Unlike orthosteric drugs, allosteric modulators can offer enhanced specificity and reduced toxicity [3, 4]. However, the structural diversity of allosteric mechanisms across protein families has long complicated efforts to systematically identify allosteric sites.

The systematic collection of experimental annotations in recent years has greatly improved the accessibility of curated data on allosteric proteins. The Allosteric Database (ASD) [5] is the most comprehensive, with derived subsets such as ASBench [6] and CASBench [7] providing filtered, benchmarking-oriented collections. AlloBench [8] was recently introduced to facilitate standardized evaluation of predictive models. The growing volume and accessibility of annotated allosteric data have enabled the development of data-driven computational approaches for allosteric site prediction, offering scalable and generalizable alternatives to experimental identification.

Computational methods for allosteric site prediction can be broadly divided into two categories. *Residue-level* methods aim to identify individual allosteric residues within a protein, typically by combining structural, evolutionary, and dynamic features at the per-residue level, as in AR-Pred [9], which couples dynamics features using elastic network models with evolutionary information, NACEN [10], which exploits residue contact energy networks, and more recently a nanoenvironment descriptor-based approach [11]. The advent of protein language models (PLMs) has further enabled *sequence-based* residue prediction without requiring a three-dimensional structure, as demonstrated by AlloFusion [12], which leverages PLM embeddings, and Kannan et al. [13], which uses PLM attention maps to identify allosteric residues directly from sequence. *Pocket (or site)-level* methods, by contrast, operate on cavities detected on the protein surface and classify them as allosteric or non-allosteric, offering a more directly actionable output for drug design. While residue-level approaches provide finer mechanistic insight, pocket-level methods are better suited for identifying druggable sites and are the focus of the present work.

Pocket-level approaches have diversified over the past decade, yet most machine learning-based models still rely on static protein conformations. Many such methods use geometry-based pocket detection tools such as *fpocket* [14] to derive physicochemical features and employ classifiers, as implemented in *Allosite* [15] and the *PASSer* family [16–18]. Other recent strategies extend these approaches by incorporating complementary structural descriptors and advanced feature selection, as in *MEF-AlloSite* [19]. Beyond static representations, several methods rely on coarse-grained dynamic analysis. *AllositePro* [20] uses normal mode analysis based on an elastic network model to evaluate protein dynamic changes upon allosteric ligand binding; *AlloPred* [21] applies normal mode perturbation analysis to capture flexibility changes, and *ESSA* (Essential Site Scanning Analysis) [22] identifies residues whose perturbation strongly alters the dispersion of global motions within elastic network-based normal modes. A related perturbation-based approach, *APOP* [23], evaluates the allosteric potential of fpocket-detected pockets by stiffening residue–residue interactions within each cavity in a Gaussian network model and ranking them based on induced shifts in global mode frequencies combined with hydrophobicity descriptors. More recently, *AllosES* [24] expanded this concept by integrating transfer entropy–based coupling derived from Gaussian network models, followed by *AlloEF* [25] and *ZHMolEReP* [26], two residue-level methods that combine dynamic and energetic features (transfer entropy and energetic frustration for the former, perturbation response scanning and free energy approximation for the latter), both requiring active site information as input.

While the coarse-grained dynamic analyses described above go beyond purely static representations, they rely on simplified physical models and a single input structure, which may not fully capture the conformational heterogeneity central to allosteric regulation. Molecular dynamics (MD) simulations offer a principled route to capture such dynamics, sampling the conformational space of a protein at atomic resolution over time. However, the computational cost of all-atom MD simulations remains substantial, particularly at the scale of large benchmark datasets. The recent emergence of generative models for conformational sampling, such as AlphaFlow [27], offers a promising and scalable alternative. More broadly, the growing interest in dynamics-aware machine learning, including methods that incorporate conformational flexibility into binding interface prediction [28], underscores the potential of integrating protein dynamics into allosteric site modelling.

In this study, we present AlloDyn, which is, to the best of our knowledge, the first dynamic-aware method for allosteric site prediction that integrates static *fpocket* descriptors with dynamic features derived from full-atom MD simulations of apo (modulator-free) protein structures, allowing direct comparison with a static-only baseline. We further evaluate AlphaFlow [27], a diffusion-based generative model that constructs conformational ensembles directly from protein sequences, as a computationally tractable alternative to MD simulations. In addition, we identify and characterize a systematic bias introduced when *fpocket* is applied to holo structures without prior removal of the allosteric modulator, leading to data leakage and artificially inflated performance estimates. AlloDyn is evaluated against several state-of-the-art allosteric site prediction methods on the D24 benchmark set [20], demonstrating competitive and consistent classification performance.

Materials and Methods

Allosteric Dataset Construction and Annotation

The dataset used in this study was prepared following a pipeline inspired by PASSerRank [18], which was developed to curate a high-quality, structurally consistent, and sequence-diverse collection of protein–modulator complexes for allosteric site prediction. Protein structures and annotations were obtained from the 2023 release of the Allosteric Database (ASD2023) [5], which contains over 2,400 experimentally validated protein–modulator entries. To ensure structural reliability, we retained only X-ray structures with resolution better than 3.0 Å, and excluded entries with incomplete allosteric sites or ambiguous modulator annotations. Pocket identification was performed using *fpocket* [14], a widely used geometry-based method. We explored two preprocessing strategies prior to pocket detection. In the first, *fpocket* was applied directly to the full structure using a flag to restrict detection to the chain annotated as allosteric in the ASD, without any prior cleaning. In the second, solvent molecules and ions were removed, the allosteric modulator was extracted from the chain, and *fpocket* was then applied to the cleaned, isolated chain. For each protein, the geometric centers (*centroids*) of the predicted pockets and the modulator were computed. The pocket with the shortest centroid-to-centroid distance to the modulator (*pocket-modulator distance*) was labeled as allosteric, and all others as non-allosteric. This modulator-based labeling strategy was adopted as residue-level allosteric site annotations are not consistently available across all entries in ASD. If the minimum distance exceeded 10 Å (*default threshold*), the entry was excluded, as such cases typically reflect failures in pocket detection [18]. Redundancy reduction was applied to obtain a sequence-diverse dataset, using

pairwise sequence identity filtering with a 30% threshold. After detection, labeling, and filtering the final dataset consisted of 353 proteins, comprising 8,847 pockets in total, of which 353 were labeled as allosteric and 8,494 as non-allosteric (Figure 1-(1)).

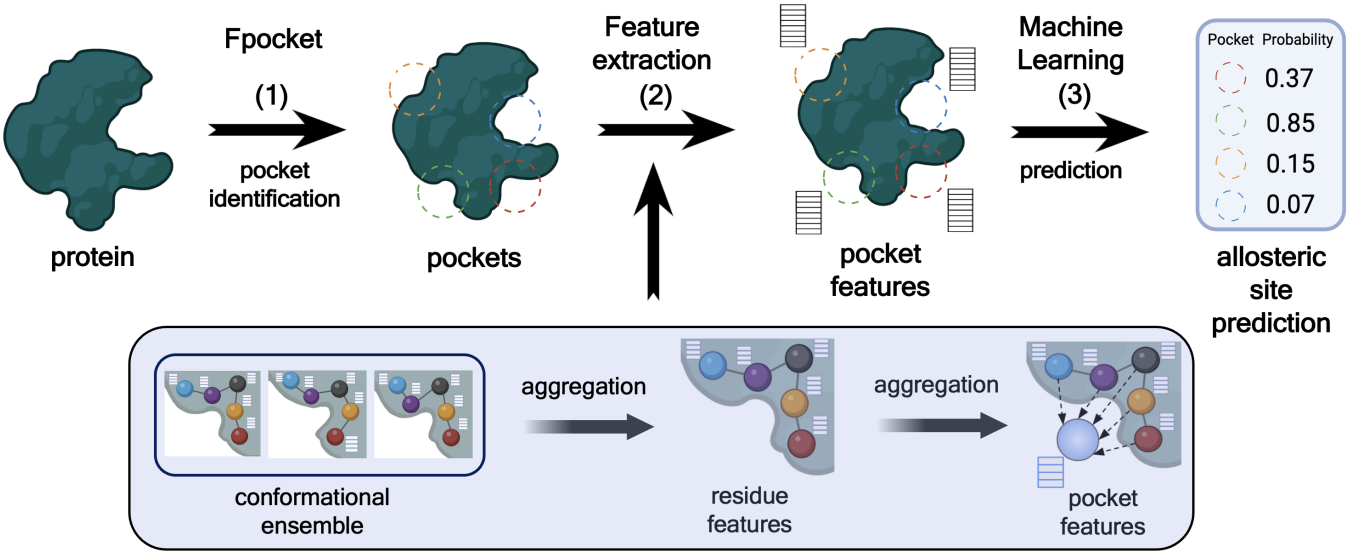


Figure 1: **Overview of the AlloDyn pipeline.** (1) *fpocket* detects pockets on the input protein structure. (2) Static features come directly from *fpocket* descriptors; dynamic features come from conformational ensembles generated by MD simulations or AlphaFlow, aggregated across the ensemble into residue-level features and then into pocket-level features. (3) XGBoost combines both feature sets and outputs an allosteric probability score for each pocket.

Generation of Conformational Ensembles

Molecular Dynamics simulations. For each of the 353 protein chains, we performed three replicates of 500-ns MD simulations, starting from their respective experimental structures. All bound ligands were removed, while ions were retained. Missing residues in regions with continuous gaps were modeled using MODELLER [29]. MD simulations were performed following the protocol described in [30], with identical simulation settings. Simulations were performed using GROMACS 2024.2 [31] with the CHARMM36m force field [32]. Systems were neutralized and adjusted to physiological ionic strength (150 mM NaCl) by adding Na^+ and Cl^- ions. Each system was solvated in a dodecahedral box of explicit TIP3P water model with a minimum buffer distance of 12 Å from the solute. Hydrogen atoms were added, and histidine protonation states were assigned using Reduce [33]. Energy minimization was carried out using the steepest descent algorithm for 5000 steps to remove steric clashes. This was followed by a multi-step equilibration phase in the NPT ensemble at 310 K, during which positional restraints were applied to protein and lipid heavy atoms and gradually released over 0.375 ns. Pressure was maintained at 1 bar using the Berendsen barostat during equilibration [34]. Production simulations were performed in the NPT ensemble with a time step of 2 fs. For each system, three independent replicates of 500 ns were generated using different initial velocities. Temperature was maintained at 310 K using the V-rescale thermostat [35], and pressure was controlled at 1 atm using the C-rescale barostat under isotropic conditions [36]. Covalent bonds involving hydrogen atoms were constrained using the LINCS algorithm [37], and long-range electrostatic interactions were treated with the Particle Mesh Ewald method [38]. Coordinates were recorded every 100 ps. For downstream analysis,

the three MD replicates were combined into a single concatenated trajectory for each protein, which served as the basis for dynamic feature extraction.

Generative models. As an alternative to classical MD simulations, we explored the use of AlphaFlow [27], a diffusion-based generative model that produces conformational ensembles directly from protein sequences. We employed its distilled version, which offers a favorable tradeoff between accuracy and computational efficiency. For consistency with MD-derived ensembles, we used the same curated protein sequences as input to AlphaFlow. For each protein entry, 25 conformations were generated. To ensure consistent atom and residue indexing, the sequences were aligned to their corresponding experimental structures prior to feature extraction.

Static and Dynamic Features Construction

Fpocket features

Each detected pocket was described using 19 descriptors computed by *fpocket*. These features capture geometric, physicochemical, and topological properties of the pocket, and include descriptors such as volume, solvent-accessible surface area (SASA), polarity, hydrophobicity, charge, and alpha-sphere statistics. Although one of these descriptors encodes residue flexibility via B-factors, all fpocket descriptors are derived from a single static structure rather than from conformational ensembles, and are referred to as *static* or *fpocket* features throughout this work.

Dynamic features

We computed dynamic descriptors from conformational ensembles to capture various aspects of pocket dynamics (Figure 1-(2)). Pockets were defined on the experimental structures using *fpocket* and represented as fixed sets of residues. Dynamic descriptors were computed by tracking these residues across conformational ensembles at the pocket level. Dynamic features were organised into three groups as described below (full list of features is available in Supplementary Table S1).

Compactness and shape features All features in this group were computed from the Cα atom coordinates of pocket residues using MDAnalysis [39]. For each frame, the pocket centroid was computed as $\mathbf{c}_t = \frac{1}{N}\sum_i \mathbf{r}_i^t$ over the N Cα atoms of the pocket. Based on this definition, we extracted nine descriptors capturing pocket geometry and variability across the ensemble. These include the standard deviation of the centroid position, as well as the mean and standard deviation of the minimum, average, and maximum Cα-to-centroid distances. In addition, mean pairwise inter-residue distances were computed per frame and summarised by their mean and standard deviation across the ensemble. Pocket compactness was quantified using the root mean square radius, $r_{\mathrm{rms}}(t) = \sqrt{\frac{1}{N}\sum_{i=1}^{N} \|\mathbf{r}_i^t - \mathbf{c}_t\|^2}$, computed per frame and summarised by its mean and standard deviation. Residue-level flexibility was captured through per-residue RMSF values, aggregated as mean and standard deviation across pocket residues. Finally, pocket shape was characterised per frame by the eigenvalues of the 3×3 spatial covariance matrix of Cα positions, $\mathbf{C}_t = \frac{1}{N-1}\sum_{i=1}^{N}(\mathbf{r}_i^t - \mathbf{c}_t)(\mathbf{r}_i^t - \mathbf{c}_t)^\top$. The resulting eigenvalues $\lambda_1 \geq \lambda_2 \geq \lambda_3$ describe the principal axes of the instantaneous spatial distribution of pocket residues, and were used to derive anisotropy $A = \lambda_3/\lambda_1$ and sphericity $S = (\lambda_1\lambda_2\lambda_3)^{1/3}/(\lambda_1+\lambda_2+\lambda_3)$, both summarised by their mean and standard deviation over the ensemble (17 features).

SASA-based features Pocket solvent-accessible surface area is computed per frame using MDTraj [40]. Both absolute (Å^2) and relative (fraction of total protein SASA) values are recorded, with mean, standard deviation, range, and coefficient of variation computed for each, leading to a total of 8 features per pocket.

Dynamic coupling features Pairwise couplings between all protein residues are quantified using four matrices. The *dynamic cross-correlation* (Pearson) measures the normalised cross-covariance of C_α positional fluctuations [41]:

$$\mathrm{DCC}_{ij} = \frac{\langle \Delta\mathbf{r}_i \cdot \Delta\mathbf{r}_j \rangle}{\sqrt{\langle |\Delta\mathbf{r}_i|^2 \rangle \langle |\Delta\mathbf{r}_j|^2 \rangle}}, \qquad \Delta\mathbf{r}_i = \mathbf{r}_i - \langle \mathbf{r}_i \rangle \tag{1}$$

Mutual information (MI) and *generalised correlation* (GC) capture both linear and nonlinear couplings [42]. Under the Gaussian approximation:

$$\mathrm{MI}_{ij} = \frac{1}{2}\left[\ln\det(\Sigma_i) + \ln\det(\Sigma_j) - \ln\det(\Sigma_{ij})\right], \tag{2}$$

where Σ_i, Σ_j are the covariance matrices of residues i and j, and Σ_{ij} is their joint covariance matrix. GC [42] is derived from MI as:

$$\mathrm{GC}_{ij} = \sqrt{1 - \exp\left(-\frac{2}{3}\,\mathrm{MI}_{ij}\right)}, \tag{3}$$

Finally, the *communication propensity* (CP) quantifies the variance of the inter-residue distance across the ensemble [43]:

$$\mathrm{CP}_{ij} = \langle (d_{ij} - \bar{d}_{ij})^2 \rangle, \tag{4}$$

where d_{ij} is the instantaneous distance between C_α atoms of residues i and j. GC, MI, and CP were computed using ComPASS [44]. For each matrix, three sets of pocket-level statistics are derived: *intra-pocket* couplings (mean, max, and std of all residue pairs within the pocket), *global connectivity* (mean and std of couplings between pocket residues and the rest of the protein), and *inter-pocket* couplings (statistics over couplings between this pocket and every other detected pocket).

Model Training and Evaluation

Classifier and hyperparameter optimization. We trained gradient-boosted decision tree models using XGBoost [45], selected for its speed, scalability, and effectiveness on imbalanced classification tasks (Figure 1-(3)). To prevent data leakage and ensure generalization across protein structures, we employed a group-aware splitting strategy: both the train/test split (80/20) and the 5-fold cross-validation for hyperparameter tuning were performed such that all pockets from the same protein were consistently assigned to a single partition. Hyperparameters were optimized via grid search, with model selection based on the average F1 score across cross-validation folds. To address the substantial class imbalance between allosteric and non-allosteric pockets, the `scale_pos_weight` parameter, which controls the weight assigned to the positive class, was included in the hyperparameter search space. The best-performing configuration was used to retrain the model on the full training set and evaluated on the independent test set.

Evaluation of dynamic feature contribution. To obtain robust performance estimates, the full training and evaluation pipeline was repeated across 50 random group-based train/test splits. This procedure was applied to the static baseline model, as well as to models augmented with dynamic features from MD

simulations or AlphaFlow-generated conformational ensembles. To ensure a fair comparison between MD- and AlphaFlow-derived features, only entries with fewer than 862 residues were retained, as this represented the upper length limit for AlphaFlow, resulting in 339 entries (7,779 pockets, non-allosteric-to-allosteric ratio ∼22:1). Performance was evaluated primarily using the F1 score, given the strong class imbalance; precision, recall, MCC, AUC-ROC, and AUC-PR were additionally computed.

Benchmarking against state-of-the-art methods. To compare our approach with existing methods, we used the D24 benchmark set from AllositePro [20]. Entries sharing high sequence identity with D24 proteins were removed from our training dataset using pairwise sequence identity filtering at a 30% threshold, yielding a filtered training set of 293 entries (6,704 pockets, class ratio ∼22:1). The final model was trained on this filtered dataset using 5-fold group-aware cross-validation to select optimal hyperparameters, followed by retraining on the full filtered set. This model was then evaluated on D24 alongside PASSer (ensemble [16], AutoML [17], and ranking-based [18] variants), AllosES [24], and APOP [23]. D24 structures were processed using the same pipeline applied to our main dataset: ligands, solvent molecules, and ions were removed, and for each entry, only the single functional chain annotated as allosteric in the AllositePro set was retained (Supplementary Table S2). The pocket closest to the bound modulator was labeled as the allosteric site, resulting in exactly one positive entry per structure. For all methods, precision, recall, F1-score, MCC, AUC-ROC, and AUC-PR were computed; for methods providing only ranking scores (APOP, PASSer-ranking), all positively ranked pockets were treated as predicted positives prior to metric computation.

Results

Modulator Presence Introduces a Systematic Bias in Fpocket Descriptors

When *fpocket* is applied directly to a holo structure without prior removal of the allosteric modulator, the resulting pocket descriptors can differ substantially from those obtained after pre-cleaning. We refer to these two strategies as *biased* (no pre-cleaning) and *unbiased* (pre-cleaned structure) throughout this work. We identified two sources of inconsistency introduced by the biased approach. First, in cases where the allosteric modulator is a modified amino acid, *fpocket* may incorporate it into the pocket definition or even construct a pocket around it, directly biasing the pocket geometry and composition. Second, and more systematically, we observed that while *fpocket* does not consider the modulator during pocket *detection*, it appears to include it during *feature calculation*. In particular, SASA values are computed over the full structure including the modulator, which artificially reduces the solvent-exposed area of the allosteric pocket in the biased case. Since the *fpocket* score and druggability score are derived from regression formulas that incorporate SASA, this propagates into a systematic upward shift of these scores for allosteric pockets in the biased dataset. This effect is clearly visible in the score distributions (Figure 2A): while non-allosteric pockets show nearly identical distributions between the biased and unbiased configurations, allosteric pockets exhibit a pronounced rightward shift in *fpocket* score and druggability score under the biased configuration. Manual inspection of individual entries (Figure 2B-C, Supplementary Table S3) confirmed that the primary differences between biased and unbiased descriptors are concentrated in *fpocket* score, druggability score, and SASA-related features, while geometric descriptors such as volume, alpha-sphere statistics, and flexibility remained largely unchanged.

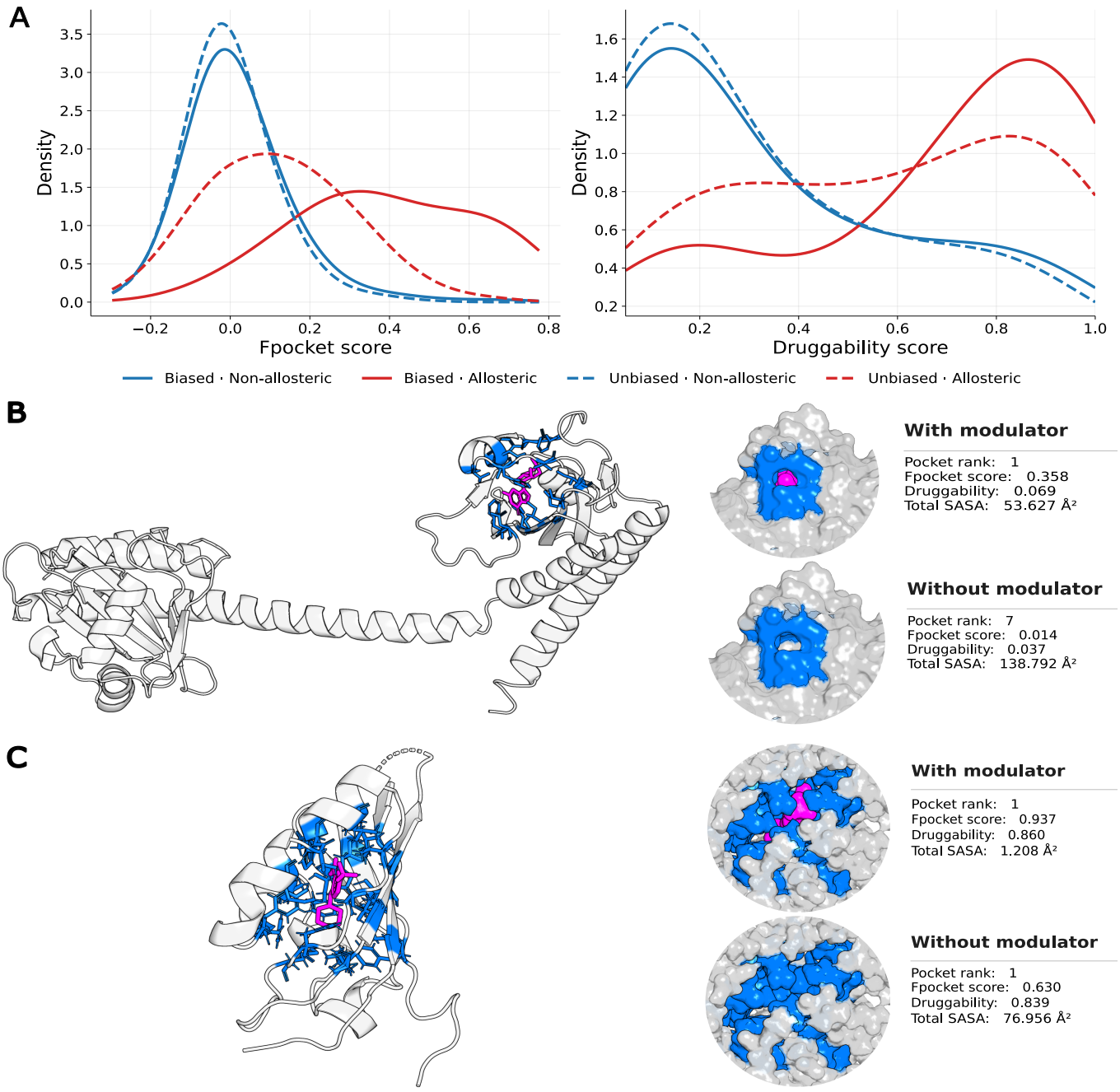


Figure 2: **Comparison of fpocket descriptors between biased and unbiased configurations. (A)** Kernel density estimates of fpocket score and druggability score across all detected pockets, split by allosteric and non-allosteric labels. **(B)** 1MC0 and **(C)** 3F1O allosteric binding sites shown in biased (modulator present) and unbiased (modulator removed) configurations. Left: full protein structure (cartoon) with allosteric residues (marine sticks) and modulator (magenta sticks). Right: zoomed surface views of the allosteric pocket (marine) with (top) and without (bottom) the modulator (magenta surface). Removal of the modulator affects the *fpocket* score (and potentially the pocket rank), druggability score, and total SASA, illustrating the systematic bias introduced when *fpocket* is applied to holo structures.

The impact of this data leakage is directly reflected in model performance (Table 1): models trained on the biased dataset achieve substantially higher F1 scores than those trained on the unbiased dataset.

However, this apparent performance gain is artifactual: feature importance analysis (Supplementary Figure S1) confirms that the *fpocket* score is the dominant predictor in the biased setting, disproportionately driving classification due to its artificially inflated values for allosteric pockets. This inflated performance does not reflect the model's ability to identify genuine allosteric features, and cannot be used as a reliable estimate of predictive performance.

Table 1: **Model performance on biased and unbiased datasets, averaged over 50 random seeds.** For each dataset, results are reported for the *fpocket*-only baseline and AlloDyn models augmented with MD- or AlphaFlow-derived (AF) dynamic features. Values are reported as mean ± standard deviation of test set performance across the 50 splits.

Source	Model	Precision	Recall	F1	MCC	AUC ROC	AUC PR
Biased	Baseline	0.57 ± 0.05	0.65 ± 0.09	0.60 ± 0.05	0.59 ± 0.05	0.94 ± 0.02	0.61 ± 0.06
	AlloDyn (Fpocket+MD)	0.54 ± 0.03	0.69 ± 0.09	0.61 ± 0.04	0.59 ± 0.05	0.95 ± 0.01	0.59 ± 0.04
	AlloDyn (Fpocket+AF)	0.56 ± 0.05	0.68 ± 0.08	0.61 ± 0.05	0.60 ± 0.05	0.95 ± 0.02	0.62 ± 0.05
Unbiased	Baseline	0.46 ± 0.06	0.41 ± 0.06	0.43 ± 0.05	0.41 ± 0.05	0.86 ± 0.03	0.40 ± 0.06
	AlloDyn (Fpocket+MD)	0.48 ± 0.06	0.42 ± 0.05	0.45 ± 0.05	0.43 ± 0.05	0.86 ± 0.02	0.41 ± 0.06
	AlloDyn (Fpocket+AF)	0.48 ± 0.06	0.42 ± 0.06	0.45 ± 0.05	0.43 ± 0.05	0.87 ± 0.02	0.40 ± 0.06

Taken together, these findings indicate that applying *fpocket* to uncleaned holo structures introduces a data leakage: allosteric pockets receive artificially inflated scores due to the presence of the modulator, making them easier to classify for reasons unrelated to genuine structural features. The unbiased preprocessing strategy, in which the modulator is removed prior to pocket detection, eliminates this artifact and provides a more reliable basis for model training and evaluation.

Dynamic Feature Augmentation Improves Allosteric Site Prediction

To evaluate the contribution of dynamic features to allosteric site prediction, we compared models trained on *fpocket* descriptors alone (baseline) to those augmented with dynamic features derived from MD simulations or AlphaFlow-generated conformational ensembles, using the unbiased dataset. Model training and evaluation were repeated across 50 random group-based train/test splits. Statistical significance of performance differences was assessed using two-tailed paired t-tests on matched splits, after confirming normality of all metric distributions with the Shapiro–Wilk test ($p > 0.05$); effect sizes were quantified using Cohen's d (more details in Supplementary Table S4).

Both AlloDyn variants significantly outperformed the static baseline across multiple metrics (Table 1). For F1 and MCC, improvements were statistically significant ($^{**}p < 0.01$) with small effect sizes ($|d| \approx 0.4$). For AUC-PR, the MD-augmented model reached a medium effect size ($|d| = 0.56$, $^{***}p < 0.001$). Models trained on MD- and AlphaFlow-derived features performed comparably across all metrics, with no statistically significant difference between the two conformational sources for F1 and MCC ($p > 0.98$, $|d| \approx 0.00$). This indicates that AlphaFlow-generated ensembles can serve as a computationally efficient alternative to MD simulations within the current classification framework.

Comparison with State-of-the-Art Methods

AlloDyn was evaluated on the D24 benchmark set against PASSer (ensemble [16], AutoML [17], and ranking-based [18] variants), AllosES [24], and APOP [23]. For this evaluation, the AlloDyn model was trained using AlphaFlow-generated conformational ensembles as the source of dynamic features. All methods were evaluated using the same set of classification metrics (Table 2). For methods providing only ranking scores (APOP, PASSer-ranking), all positively ranked pockets were treated as predicted positives when computing classification metrics. The number of evaluated pockets (N) varies across methods. AlloDyn and PASSer operate on the same pocket set ($N = 645$). The small discrepancy for APOP ($N = 641$) arises from its mandatory removal of non-standard amino acids during preprocessing, which introduces gaps in the sequence and alters the alpha-sphere clustering step in *fpocket*, resulting in slightly different pocket definitions for a small number of entries. For AllosES, the use of an older version of *fpocket* in combination with non-standard amino acid removal led to a substantially different pocket set ($N = 305$).

Table 2: **Performance comparison of allosteric site prediction methods on the D24 benchmark.** Metrics are reported at the pocket level. N denotes the number of evaluated pockets. Best values per column are shown in bold.

Method	Precision	Recall	F1	MCC	AUC-ROC	AUC-PR	N
AlloDyn	0.52	0.50	**0.51**	**0.49**	**0.93**	0.44	645
APOP	0.09	**0.95**	0.17	0.23	**0.93**	0.52	641
AllosES	0.24	0.79	0.37	0.37	0.86	**0.54**	305
AllosES*	**0.60**	0.27	0.38	0.39	0.90	0.41	645
PASSer-automl	0.50	0.27	0.35	0.35	0.91	0.32	645
PASSer-ensemble	0.45	0.23	0.30	0.31	0.86	0.38	645
PASSer-rank	0.21	0.14	0.17	0.15	0.84	0.28	645

*Re-evaluated with fpocket v.4 to match the pocket set used by all other methods ($N = 645$).

Figure 3 illustrates representative prediction outcomes on the D24 benchmark. Panel **A** shows a prediction example for 4OYA_A, human soluble adenylate cyclase, demonstrating the discriminative power of AlloDyn: while AllosES and PASSer variants assign the correct allosteric pocket a lower score and instead prioritize a distant, non-allosteric cavity, AlloDyn correctly identifies it as the top prediction. Panels **B**–**C** illustrate the fundamental limitation imposed by the pocket detection step: for 5J94_A and 4NHV_A, no *fpocket*-detected pocket falls within the 10 Å labelling threshold of the allosteric modulator, meaning all pockets were labelled as non-allosteric and the correct site could not be recovered by any method. Detailed predictions for all D24 entries across all compared methods are provided in Supplementary Figures S2–S4.

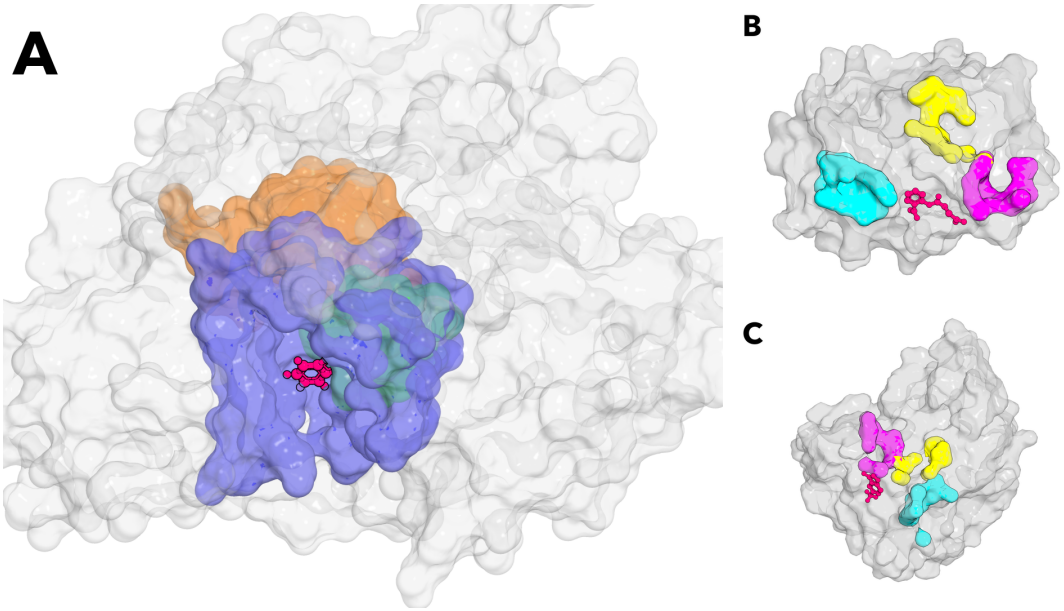


Figure 3: **Representative allosteric pocket predictions on the D24 benchmark. (A)** Prediction comparison for 4OYA_A. The correct allosteric pocket is ranked first by AlloDyn (green, buried pocket, 2.5 Å from modulator) and APOP (blue, larger binding-site representation). The difference in pocket extent between the two methods arises from APOP's removal of modified residues during preprocessing. AllosES and PASSer variants assign a lower rank to the correct allosteric pocket, instead prioritizing a distant pocket (orange, 15.3 Å from the modulator) as their top prediction. Modulator is shown in pink (sticks). **(B–C)** Failure cases for 5J94_A and 4NHV_A, respectively. No *fpocket*-detected pocket falls within the 10 Å labelling threshold of the allosteric modulator (pink sticks), showing that prediction performance is fundamentally constrained by the pocket detection step in these cases. The three closest detected pockets are shown: the nearest (magenta), second nearest (cyan), and third nearest (yellow). Protein surface is shown in light gray.

Among all evaluated methods, AlloDyn achieves the best balance between precision and recall, yielding the highest F1 score (0.51) and MCC (0.49, Table 2). APOP attains the highest recall (0.95) together with a competitive AUC-PR (0.52), but at the cost of very low precision (0.09), indicating that a large fraction of detected pockets are predicted as allosteric, leading to numerous false positives. Similarly, AllosES achieves the highest AUC-PR (0.54), suggesting strong ranking performance across thresholds, but its lower precision results in reduced F1 and MCC scores compared to AlloDyn. PASSer variants show moderate and consistent performance across metrics, with PASSer-AutoML performing best within the family. These results indicate that while APOP and AllosES are effective at assigning elevated scores to allosteric pockets, they tend to overpredict positive sites. In contrast, AlloDyn provides a more balanced tradeoff between sensitivity and specificity, leading to more reliable threshold-based classification performance. In practical applications, where experimental validation of predicted allosteric sites is costly, reducing false positives while maintaining competitive recall may represent a more useful operating regime.

Discussion

This study investigated whether dynamic features derived from conformational ensembles can improve allosteric site prediction beyond what is achievable with static *fpocket* descriptors alone. Using a curated

dataset of annotated allosteric and non-allosteric pockets, we trained and evaluated XGBoost classifiers on static features, conformational ensemble-derived dynamic features (MD or AlphaFlow-generated). Our results show that dynamic feature augmentation significantly improves classification performance when structures are properly preprocessed, while also revealing several methodological bottlenecks that limit the utility of dynamic descriptors in the current framework.

Pocket detection constitutes the first and most critical step in our pipeline, and its accuracy directly determines the upper bound of downstream prediction performance. When *fpocket* is applied to holo structures without prior removal of the allosteric modulator, SASA-based descriptors are computed over the full structure including the modulator, artificially inflating the *fpocket* score and druggability score for allosteric pockets. This constitutes a data leakage that leads to overoptimistic performance estimates. Removing the modulator prior to pocket detection eliminates this artifact and provides a more reliable basis for model training. Beyond preprocessing, *fpocket* may simply fail to detect the correct allosteric pocket in some cases, as illustrated by the two D24 entries (5J94_A and 4NHV_A) where no detected pocket fell within the labelling threshold of the modulator. In such cases, no downstream method can recover the correct site, highlighting the fundamental dependency of our approach on pocket detection quality.

Despite the dominance of static descriptors, dynamic feature augmentation led to a statistically significant improvement in F1 score on the unbiased dataset. Notably, some static *fpocket* descriptors (in particular the *fpocket* score and druggability score) remained among the top predictors across all settings, reflecting their strong discriminative power even in the absence of dynamic information. These results suggest that dynamic features are most informative when the allosteric pocket is accurately identified, highlighting the dependency of dynamic descriptors on the quality of the initial pocket definition.

Dynamic features were computed by anchoring to the set of pocket-forming residues identified in the experimental structure. While this provides a consistent reference across conformations, it is an approximation: the true pocket boundary may shift during simulation, and transient or cryptic regions may not be captured by residues defined on a single static structure. More sophisticated pocket tracking strategies, such as ensemble-based pocket detection or adaptive residue mapping across conformations, could improve the accuracy of dynamic descriptors. Furthermore, obtaining pocket-level features requires a double aggregation - from the conformational ensemble to per-residue features, and then from residues to the pocket level - which inevitably discards spatial and temporal resolution. More expressive architectures, such as graph neural networks with learnable aggregation, could better preserve this information and more fully exploit the rich structural information contained in conformational ensembles.

Correlation-based features capture residue-residue communication patterns that are conceptually well-suited to allosteric regulation. However, their computation requires knowledge of which residues belong to each pocket, and the current approach uses all detected pockets for inter-pocket correlation calculations due to the absence of explicit orthosteric site annotations in the dataset. This introduces noise in the inter-pocket and global connectivity features, as non-functional pockets are treated on equal footing with biologically relevant ones. Future work incorporating orthosteric site annotations would allow more targeted and interpretable correlation descriptors.

Despite known structural imperfections such as occasional atomic clashes, models trained on AlphaFlow-derived dynamic features achieved performance comparable to those using MD-derived ensembles, with no statistically significant differences. A current limitation of AlphaFlow is that it operates on single protein chains, whereas MD simulations can accommodate multimeric complexes. For proteins whose allosteric regulation depends on inter-chain interactions, this represents a meaningful constraint. Nevertheless, for monomeric systems, AlphaFlow offers a practical and computationally efficient alternative to MD for conformational sampling in this context.

Overall, our results demonstrate that integrating dynamic features into allosteric site prediction is a promising direction, but its effectiveness is conditioned on reliable pocket detection, accurate structural annotations, and sufficiently expressive feature representations. Expanding annotated datasets with orthosteric site information and developing architectures capable of learning directly from raw conformational ensembles are key directions for future work.

Data and code availability

The source code and training datasets are publicly available on GitLab at https://gitlab.inria.fr/vpryakhi/AlloDyn.

Acknowledgements

YK was supported by the French National Research Agency (ANR) under the France 2030 grant reference number ANR-24-RRII-0002 operated by the Inria Quadrant Program. Computations were performed using resources from the Grid'5000 and MBI computing platforms. This work was granted access to the HPC resources of IDRIS under the allocation 2025-A0180714660 granted to Y.K. made by GENCI.

Conflicts of Interest

The authors declare no conflicts of interest.

References

1. Nina M. Goodey and Stephen J. Benkovic. Allosteric regulation and catalysis emerge via a common route. *Nat. Chem. Biol.*, 4(8):474–482, 2008.

2. Mingyu Li, Xiaobin Lan, Xun Lu, and Jian Zhang. A Structure-Based Allosteric Modulator Design Paradigm. *Health Data Sci.*, 3:0094, 2023.

3. Ashok Kumar Grover. Use of Allosteric Targets in the Discovery of Safer Drugs. *Med. Princ. Pract.*, 22(5):418–426, 2013.

4. Nan Wu, Léonie Strömich, and Sophia N. Yaliraki. Prediction of allosteric sites and signaling: Insights from benchmarking datasets. *Patterns*, 3(1):100408, 2022.

5. Jixiao He, Xinyi Liu, Chunhao Zhu, Jinyin Zha, Qian Li, Mingzhu Zhao, Jiacheng Wei, Mingyu Li, Chengwei Wu, Junyuan Wang, Yonglai Jiao, Shaobo Ning, Jiamin Zhou, Yue Hong, Yonghui Liu, Hongxi He, Mingyang Zhang, Feiying Chen, Yanxiu Li, Xinheng He, Jing Wu, Shaoyong Lu, Kun Song, Xuefeng Lu, and Jian Zhang. ASD2023: Towards the integrating landscapes of allosteric knowledgebase. *Nucleic Acids Res.*, 52(D1):D376–D383, 2024.

6. Wenkang Huang, Guanqiao Wang, Qiancheng Shen, Xinyi Liu, Shaoyong Lu, Lv Geng, Zhimin Huang, and Jian Zhang. ASBench: Benchmarking sets for allosteric discovery. *Bioinformatics*, 31(15):2598–2600, 2015.

7. A Zlobin, D Suplatov, K Kopylov, and V Švedas. Casbench: a benchmarking set of proteins with annotated catalytic and allosteric sites in their structures. *Acta Naturae*, 11(1 (40)):74–80, 2019.

8. Dibyajyoti Maity and Baofu Qiao. AlloBench: A Data Set Pipeline for the Development and Benchmarking of Allosteric Site Prediction Tools. *ACS Omega*, 10(17):17973–17982, 2025.

9. Sambit K. Mishra, Gaurav Kandoi, and Robert L. Jernigan. Coupling dynamics and evolutionary information with structure to identify protein regulatory and functional binding sites. *Proteins*, 87(10):850–868, 2019.

10. Wenying Yan, Guang Hu, Zhongjie Liang, Jianhong Zhou, Yang Yang, Jiajia Chen, and Bairong Shen. Node-Weighted Amino Acid Network Strategy for Characterization and Identification of Protein Functional Residues. *J. Chem. Inf. Model.*, 58(9):2024–2032, 2018.

11. Folorunsho Bright Omage, José Augusto Salim, Ivan Mazoni, Inácio Henrique Yano, Luiz Borro, Jorge Enrique Hernández Gonzalez, Fabio Rogerio De Moraes, Poliana Fernanda Giachetto, Ljubica Tasic, Raghuvir Krishnaswamy Arni, and Goran Neshich. Protein allosteric site identification using machine learning and per amino acid residue reported internal protein nanoenvironment descriptors. *Comput. Struct. Biotechnol. J.*, 23:3907–3919, 2024.

12. Jiabin Huang, Dongliang Guo, Yapeng Liu, Yanfen Wang, and Mengya Lv. Allofusion: Allosteric Site Prediction Based on Language Models and Multi-Feature Fusion. *J. Chem. Inf. Model.*, 65(16):8858–8870, 2025.

13. Gokul R. Kannan, Brian L. Hie, and Peter S. Kim. Single-sequence, structure free allosteric residue prediction with protein language models. *bioRxiv*, 2024.

14. Vincent Le Guilloux, Peter Schmidtke, and Pierre Tuffery. Fpocket: An open source platform for ligand pocket detection. *BMC Bioinform.*, 10(1):168, 2009.

15. Wenkang Huang, Shaoyong Lu, Zhimin Huang, Xinyi Liu, Linkai Mou, Yu Luo, Yanlong Zhao, Yaqin Liu, Zhongjie Chen, Tingjun Hou, and Jian Zhang. Allosite: A method for predicting allosteric sites. *Bioinformatics*, 29(18):2357–2359, 2013.

16. Hao Tian, Xi Jiang, and Peng Tao. PASSer: Prediction of allosteric sites server. *Mach. Learn.: Sci. Technol.*, 2(3):035015, 2021.

17. Sian Xiao, Hao Tian, and Peng Tao. PASSer2.0: Accurate Prediction of Protein Allosteric Sites Through Automated Machine Learning. *Front. Mol. Biosci.*, 9:879251, 2022.

18. Hao Tian, Sian Xiao, Xi Jiang, and Peng Tao. PASSerRank: Prediction of allosteric sites with learning to rank. *J. Comput. Chem.*, 44(28):2223–2229, 2023.

19. Sadettin Y. Ugurlu, David McDonald, and Shan He. MEF-AlloSite: An accurate and robust Multimodel Ensemble Feature selection for the Allosteric Site identification model. *J. Cheminform.*, 16(1):116, 2024.

20. Kun Song, Xinyi Liu, Wenkang Huang, Shaoyong Lu, Qiancheng Shen, Lu Zhang, and Jian Zhang. Improved Method for the Identification and Validation of Allosteric Sites. *J. Chem. Inf. Model.*, 57(9):2358–2363, 2017.

21. Joe G Greener and Michael Je Sternberg. AlloPred: Prediction of allosteric pockets on proteins using normal mode perturbation analysis. *BMC Bioinform.*, 16(1):335, 2015.

22. Burak T. Kaynak, Ivet Bahar, and Pemra Doruker. Essential site scanning analysis: A new approach for detecting sites that modulate the dispersion of protein global motions. *Comput. Struct. Biotechnol. J.*, 18:1577–1586, 2020.

23. Ambuj Kumar, Burak T. Kaynak, Karin S Dorman, Pemra Doruker, and Robert L. Jernigan. Predicting allosteric pockets in protein biological assemblages. *Bioinformatics*, 39(5):btad275, 2023.

24. Fangrui Hu, Fubin Chang, Lianci Tao, Xiaohan Sun, Lamei Liu, Yingchun Zhao, Zhongjie Han, and Chunhua Li. Prediction of Protein Allosteric Sites with Transfer Entropy and Spatial Neighbor-Based Evolutionary Information Learned by an Ensemble Model. *J. Chem. Inf. Model.*, 64(15):6197–6204, August 2024.

25. Jilong Zhang, Xiaohan Sun, Zhixiang Wu, Jingjie Su, Xinyu Zhang, and Chunhua Li. AlloEF: An Ensemble Model for Protein Allosteric Site Identification Based on Transfer Entropy and Energetic Frustration. *J. Phys. Chem. B*, 130(19):4970–4981, 2026.

26. Xing Ke, Haoquan Liu, Jian Wang, Wenxin Guo, He Feng, and Yunjie Zhao. ZHMolEReP: An Energy Response Strategy for Protein Allosteric Site Prediction. *J. Chem. Inf. Model.*, 2026.

27. Bowen Jing, Bonnie Berger, and Tommi Jaakkola. AlphaFold meets flow matching for generating protein ensembles. *arXiv preprint arXiv:2402.04845*, 2024.

28. Omid Mokhtari, Sergei Grudinin, Yasaman Karami, and Hamed Khakzad. DynamicGT: A dynamic-aware geometric transformer model to predict protein-binding interfaces in flexible and disordered regions. *Cell Syst.*, 17(1), 2026.

29. Andrej Šali and Tom L. Blundell. Comparative protein modelling by satisfaction of spatial restraints. *Journal of molecular biology*, 234(3):779–815, 1993.

30. Omid Mokhtari, Emmanuelle Bignon, Hamed Khakzad, and Yasaman Karami. DynaRepo: The repository of macromolecular conformational dynamics. *Nucleic Acids Res.*, 54(D1):D393–D401, 2026.

31. David Van Der Spoel, Erik Lindahl, Berk Hess, Gerrit Groenhof, Alan E. Mark, and Herman J. C. Berendsen. GROMACS: Fast, flexible, and free. *J. Comput. Chem.*, 26(16):1701–1718, 2005.

32. Jing Huang, Sarah Rauscher, Grzegorz Nawrocki, Ting Ran, Michael Feig, Bert L De Groot, Helmut Grubmüller, and Alexander D MacKerell. CHARMM36m: An improved force field for folded and intrinsically disordered proteins. *Nat. Methods*, 14(1):71–73, 2017.

33. J.Michael Word, Simon C. Lovell, Jane S. Richardson, and David C. Richardson. Asparagine and glutamine: Using hydrogen atom contacts in the choice of side-chain amide orientation. *J. Mol. Biol.*, 285(4):1735–1747, 1999.

34. Herman J. C. Berendsen, Johan P. M. Postma, Wilfred F. van Gunsteren, A. Dinola, and Jan R. Haak. Molecular dynamics with coupling to an external bath. *J. Chem. Phys.*, 81:3684–3690, 1984.

35. Giovanni Bussi, Davide Donadio, and Michele Parrinello. Canonical sampling through velocity rescaling. *J. Chem. Phys.*, 126(1):014101, 2007.

36. Mattia Bernetti and Giovanni Bussi. Pressure control using stochastic cell rescaling. *J. Chem. Phys.*, 153(11):114107, 2020.

37. Berk Hess, Henk Bekker, Herman J. C. Berendsen, and Johannes G. E. M. Fraaije. LINCS: A linear constraint solver for molecular simulations. *J. Comput. Chem.*, 18(12):1463–1472, 1997.

38. Darrin M. York, Tom A. Darden, and Lee G. Pedersen. The effect of long-range electrostatic interactions in simulations of macromolecular crystals: A comparison of the Ewald and truncated list methods. *J. Chem. Phys.*, 99(10):8345–8348, 1993.

39. Richard Gowers, Max Linke, Jonathan Barnoud, Tyler Reddy, Manuel Melo, Sean Seyler, Jan Domański, David Dotson, Sébastien Buchoux, Ian Kenney, and Oliver Beckstein. MDAnalysis: A Python Package for the Rapid Analysis of Molecular Dynamics Simulations. In *Python in Science Conference*, pages 98–105, Austin, Texas, 2016.

40. Robert T. McGibbon, Kyle A. Beauchamp, Matthew P. Harrigan, Christoph Klein, Jason M. Swails, Carlos X. Hernández, Christian R. Schwantes, Lee-Ping Wang, Thomas J. Lane, and Vijay S. Pande. MDTraj: A Modern Open Library for the Analysis of Molecular Dynamics Trajectories. *Biophys. J.*, 109(8):1528–1532, 2015.

41. Toshiko Ichiye and Martin Karplus. Collective motions in proteins: A covariance analysis of atomic fluctuations in molecular dynamics and normal mode simulations. *Proteins*, 11(3):205–217, 1991.

42. Oliver F. Lange and Helmut Grubmüller. Generalized correlation for biomolecular dynamics. *Proteins*, 62(4):1053–1061, 2006.

43. Yasaman Karami, Elodie Laine, and Alessandra Carbone. Dissecting protein architecture with communication blocks and communicating segment pairs. *BMC Bioinform.*, 17(S2):S13, 2016.

44. Sneha Bheemireddy, Roy González-Alemán, Emmanuelle Bignon, and Yasaman Karami. Communication Pathway Analysis within Protein-Nucleic Acid Complexes. *J. Chem. Theory Comput.*, 21(17):8255–8266, 2025.

45. Tianqi Chen and Carlos Guestrin. XGBoost: A Scalable Tree Boosting System. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 785–794, 2016.

DrMab : Framework to Track Mutations of concern in Respiratory Viruses

Jérôme BOURRET[1,2,3], Marie-Anne RAMEIX-WELTI[1,3], and Frédéric LEMOINE[1,2,3]

1 Institut Pasteur, National Reference Center for Respiratory Viruses, Université Paris Cité, Paris F-75015, France

2 Institut Pasteur, Bioinformatics and Biostatistics Hub, Université Paris Cité, Paris F-75015, France

3 Institut Pasteur, Molecular Mechanisms of Multiplication of Pneumovirus, Université Versailles St-Quentin en Yvelines, Paris-Saclay INSERM UMR 1173 (2I), Assistance Publique des Hôpitaux de Paris, Paris, France

Corresponding Author: jerome.bourret@pasteur.fr

Keywords

Drug resistance, respiratory viruses, surveillance, epidemics, mutations, workflows.

1. Introduction

Respiratory viruses such as Influenza viruses and Respiratory Syncytial Virus (RSV) have long been major causes of hospitalizations due to acute respiratory infections, with SARS-CoV-2 emerging as a major global health threat in recent years. Tracking the emergence, evolution, and spread of these viruses is crucial for anticipating epidemic progression and implementing measures to mitigate public health impacts.

Recent progress in genomic surveillance has enabled the establishment of highly robust systems for monitoring epidemics. For instance, full-genome viral sequencing facilitates near real-time tracking of viral evolution through the construction and continuous updating of phylogenetic trees. These capabilities further support the development of standardized lineage nomenclatures for tracking viral clades of concern, detecting evolutionary events such as recombination and reassortment (in segmented viruses), and monitoring the emergence of concerning mutations.

These mutations of concern typically fall into several categories: host-adaptive mutations (e.g., facilitating transmission of influenza viruses from avian to human hosts), immune evasion mutations, or antiviral resistance mutations (conferring reduced susceptibility to therapeutic agents, often known as Drug Resistance Mutations or DRMs). Surveillance systems must enable the rapid and accurate detection of these mutations through routine genomic sequencing to facilitate timely public health response on an individual (e.g. for custom treatments) or a global scale (e.g. surveillance of mutational profile in the population). Crucially, monitoring must capture these variants whether they dominate viral populations or remain at low frequencies, as both contexts inform predictions regarding treatment outcomes and evolutionary trends.

In this context, bioinformatics tools play a crucial role. They must i) accurately analyze complete genome sequences from routine surveillance data (and integrate themselves easily in already existing workflows), ii) detect early indicators such as low-frequency mutations, and iii) rely on complete, flexible, and regularly updated mutation databases.

Several bioinformatics tools have been developed to identify mutations of concern, typically by aligning query sequences to reference genomes and matching the detected variations against curated catalogs of known mutations. However, most of these approaches fail to satisfy every requirement outlined above. Specifically, tools like `Flu/RSVSurver` (FluSurver, n.d.), `Sierra` (Tzou et al., 2022), and `BMA` (Salvatierra & Florez, 2016) operate exclusively on consensus sequences, precluding the analysis of low-frequency minor (intra-host) variants. Meanwhile, specialized tools including `SABRes` (Fong et al., 2023), `HerpesDRG` (Charles et al., 2024), and `DR_SEQUAN` (Garriga & Menéndez-Arias, 2006) are limited to specific viral species with restricted configurability.

To overcome these limitations, we developed DrMab, a comprehensive and flexible framework that analyzes viral whole-genome viral data (already assembled genomes or raw reads) to identify mutations of concern at the viral population level. To do so, it leverages a curated database of mutations or a user provided list of mutations.

2. DrMab framework

DrMab functionalities

DrMab comprises two core modules. The first is a manually curated and continuously updated database of mutations of concern for SARS-CoV-2 (based on the Cov-RDB database; Tzou et al., 2022), influenza viruses (based on the WHO guidelines ; WHO, n.d.), several avian influenza (list of mutations curated by experts and based on literature review) and RSV (based on the Virus French Resistance database ; Virus French Database, n.d.), reflecting current knowledge regarding variant functional impacts. In total, 806 mutations for 18 strains of viruses are recorded on our database. The second module consists of a Nextflow-based computational workflow that processes FASTQ, FASTA, and VCF inputs alongside configurable resources, including the mutation database and corresponding reference genomes.

The workflow is designed to limit manual intervention, and executes the following steps, depending on the kind of inputs: For FASTQ files, the pipeline first selects the most appropriate reference genome for each sample and maps the reads with minimap2 (Li, 2021). It then executes two concurrent variant-calling procedures: one to detect minor variants (intra-host variants) and another to generate a consensus sequence. The consensus - generated with iVar (Grubaugh et al., 2019) - is aligned with MAFFT (Katoh et al., 2019) to the reference coordinate system (against which database mutations are annotated) to identify major variants for comparison against the curated database. Simultaneously, minor variant coordinates - estimated with iVar - are converted to this standardized reference framework prior to database screening. When VCF files are provided, the workflow bypasses alignment and variant calling, proceeding directly to coordinate standardization and extraction of mutations of concern. In all cases, nucleotide sequences of all ORFs as defined in a reference gff annotation file (even when they are overlapping or subjected to frameshift) are translated into amino acids. This is done using a dedicated tool for minor variants and with goalign (Lemoine & Gascuel, 2021) for consensus sequences, and accounts for multiple minor variants occurring within individual codons and appropriately handles degenerate nucleotides codes. Figure 1 illustrates the simplified FASTQ processing subworkflow.

The DrMab workflow generates two primary outputs:

- A list of minor variants of concern: these are low frequency mutations
- A list of dominant variants of concern: these are present in the consensus sequence

Both types are classified as:

- "True" (the specific mutation of concern),
- "Positional" (mutations at the same site as a known variant but with a different amino acid), or
- "Potential" (inferred from degenerated codons, applicable only to dominant variants).

In these outputs, each entry includes additional descriptive metadata such as:

- The type of mutation (DRM, vaccine-induced immunity escape, host-adaptation), as described on the curated databases.
- The variant's predicted functional consequences (level of reduced sensitivity, modifications of cellular tropism, premature stop codon potentially making the protein non-functional, etc.).

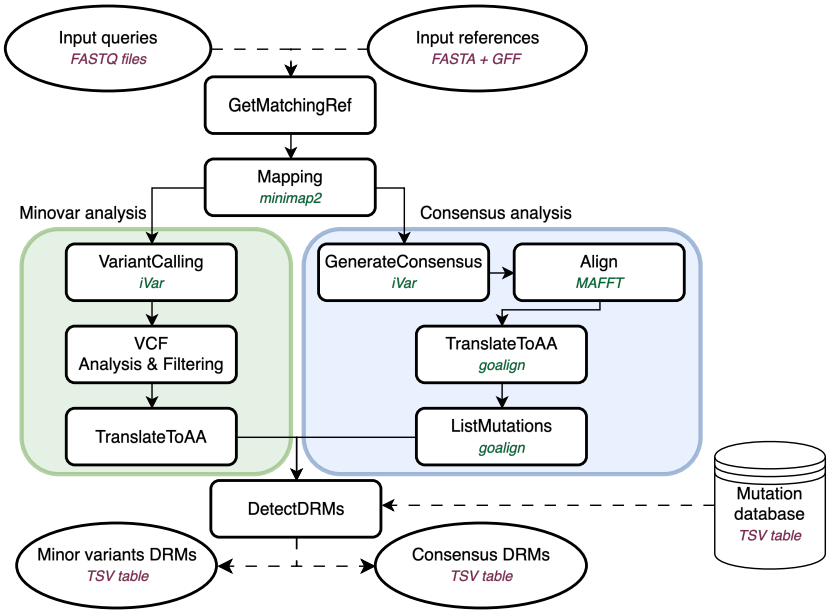


Figure 1 `DrMab` FASTQ Workflow.

`DrMab` implementation

The `DrMab` workflow is developed using Nextflow, a choice that guarantees reproducibility, scalability, and component reusability. By isolating each analytical step within dedicated software environments (containers), the pipeline ensures straightforward setup, deployment, and maintenance.

Designed with modularity in mind, `DrMab` offers distinct subworkflows tailored to specific input types (FASTA, FASTQ, and VCF) and includes dedicated pipelines for analyzing detected DRMs. It provides specialized analytical branches for RSV, SARS-CoV-2, and Influenza datasets, all of which utilize a common foundation of base modules to ensure consistency and reusability across workflows.

`DrMab` is publicly available at https://gitlab.pasteur.fr/cnrvir/pipelines/drmab/.

Together, these features establish `DrMab` as a versatile, modular platform for variant-of-concern analysis in the context of epidemic monitoring.

3. Use cases

`DrMab` is currently used by the National Reference Center for Respiratory Viruses to detect DRMs and other concerning mutations in influenza, RSV, and SARS-CoV-2 samples collected in France. We evaluated `DrMab` on two distinct datasets: 1) 386 French H1N1pdm samples collected during the 2024-2025 epidemic season, and 2) 858 French RSV samples from the 2025 POLYRES Cohort (Fourati et al., 2026). We finally compared the results obtained with DrMab and FLU/RSVSurver.

Analysis of the H1N1pdm dataset

The H1N1pdm dataset revealed rare and sporadic emergences of resistance mutations, with no clear spatial nor temporal pattern (Table 1). Most of these mutations (except H275Y) confer only mild or low resistance to treatments targeting NA or PA segments. Our findings show no evidence of fixation of these mutations, indicating that such mutational profiles were likely outcompeted and eliminated from viral populations due to fitness costs.

Table 1 H1N1pdm influenza samples from France's 2024-2025 season with detected DRMs in the NA or PA segments. The "Resistance" column specifies the affected antiviral treatment and its resistance level (in parentheses), while the "Mutation" column shows the frequency of each mutation in sequencing reads (as a percentage). Sample origins are listed by French administrative region. The bold line indicates the minor variant not detected by FLUSurver (written as H275X instead of H275Y).

ID (EPI_ISL)	Mutation (frequency)	Resistance against (level)	Location	Date
19850563	**NA:H275Y(0.46)**	**Oseltamivir, Peramivir(strong), zanamivir, laninamivir(low)**	**Basse-Normandie**	**2025-03-21**

19850562	NA:H275Y(1)	Oseltamivir, Peramivir (strong), zanamivir, laninamivir (low)	Basse-Normandie	2025-02-21
19669923	NA:I223T(1)	Oseltamivir (mild), Zanamivir, Peramivir, Laninamivir (low)	Île-de-France	2024-12-20
19690751	PA:I38V(1)	Baloxavir (low)	Nord-Pas-de-Calais	2024-12-24
19554791	NA:S247N(1)	Oseltamivir, Zanamivir, Peramivir, Laninamivir (low)	Alsace	2024-10-29
19655854	NA:S247N(1)	Oseltamivir, Zanamivir, Peramivir, Laninamivir (low)	Alsace	2024-12-02
19669949	NA:S247N(1)	Oseltamivir, Zanamivir, Peramivir, Laninamivir (low)	Nord-Pas-de-Calais	2024-12-17

Analysis of the RSV 2025 dataset

Regarding the RSV dataset, we aimed to evaluate the introduction of Nirsevimab in France by comparing mutational profiles between breakthrough samples (treated infants that are nonetheless infected by the virus ; n = 419) and untreated control samples (untreated infants ; n = 439) during the 2024–2025 epidemic season. Among breakthrough samples, `DrMab` identified DRMs in 2 RSV-A and 23 RSV-B samples, including 7 minor variants present at low frequencies (<50%). The lack of evolutionary clustering in these resistance patterns (no clear grouping by clade) suggests sporadic emergence (convergence) rather than lineage-driven propagation (Figure 2). Intriguingly, only one DRM was detected nearly twelve months post-treatment, indicating that while most resistance mutations are rapidly eliminated in the absence of continued selective pressure, rare instances of persistence may occur without achieving broader epidemic spread (Fourati et al., 2026).

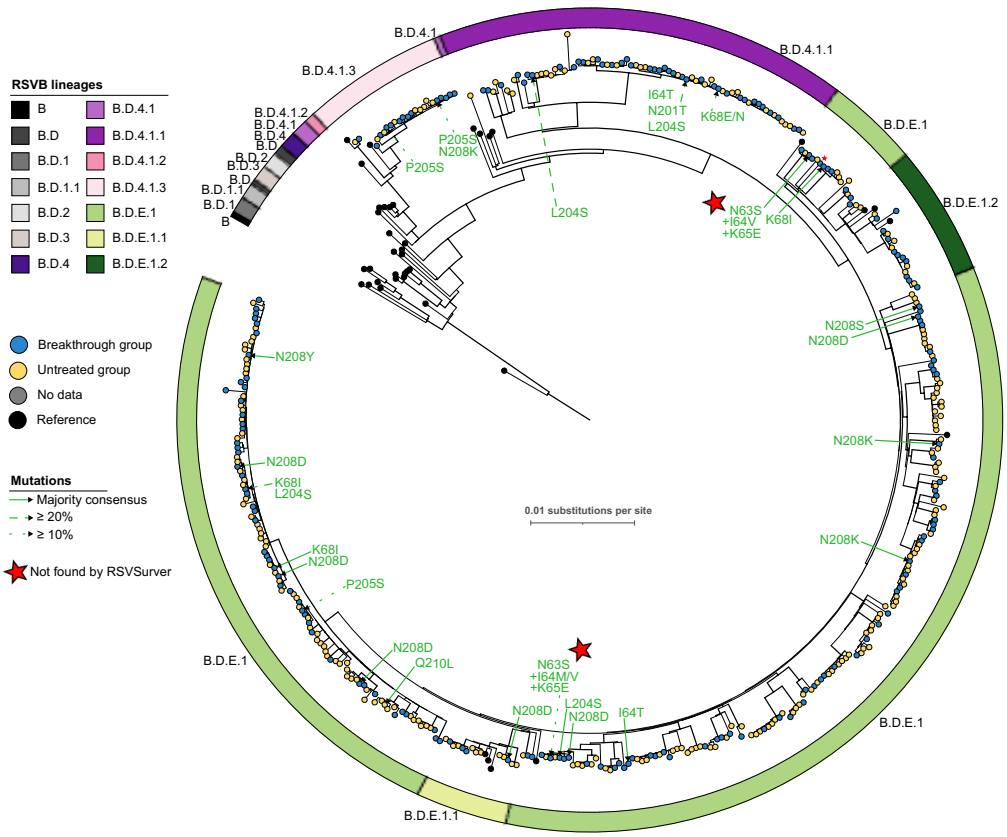


Figure 2 RSVB Phylogeny of RSVB samples from the POLYRES study. This phylogeny has been inferred from the full-length genome sequences with a Maximum-Likelihood approach (GTR model, bootstrap of 1000). Tip color indicates the breakthrough (blue), unexposed (yellow), or external sequences used as a context for this analysis (black). The colored ribbon around the tree indicates the RSV lineages to which the tips belong. Branches with a bootstrap value of 90% or higher are highlighted with a thicker line. Green lines pinpoint individuals with mutations of interest on epitope Φ of the F protein. Green dashed lines pinpoint individuals where those mutations are found in minor populations (>10% and >20% as represented in the figure). Resistance-associated substitutions co-occurring in the same viral population are separated with a « + » sign. Red stars show the co-occurring mutations not detected by RSVSurver. This phylogeny has been inferred with `IQ-TREE` (v2.2.5). The full version of this Figure can be found in Fourati et al., study (2026).

4. Conclusion

Altogether, these results show that `DrMab` is a robust and adaptable framework designed for routine and large-scale surveillance of clinically relevant mutations in major respiratory viruses, as well as for sporadic viral population / cohort analyses. `DrMab` requires few inputs, with only a predefined list of concerning mutations and a reference genome that can be both customized by the user, making it a generic framework adaptable to many viruses. Our validation across two datasets demonstrated its accuracy in detecting key mutations.

The National Reference Center for Respiratory Viruses now routinely uses `DrMab` to detect DRMs in influenza, RSV, and SARS-CoV-2 samples from France. Beyond standard surveillance, the tool is also applied in targeted scenarios, such as tracking resistance in treated patients.

Future development will focus on identifying co-occurring mutations within viral populations through haplotype estimation, as certain influenza DRMs exhibit enhanced resistance when combined (e.g., with tools such as `Haploflow`, `Virus-VG` or `SAVAGE` (Baaijens et al., 2017, 2019; Fritz et al., 2021). Additionally, we plan to include phylogenetic placements of query sequences on reference trees, enabling evolutionary interpretations of the detected variants. Finally, we plan to extend `DrMab` to support other viruses through updates of the database. `DrMab` is available as a standalone workflow for local or high-performance computing environments and will soon be accessible via a dedicated web platform.

References

Baaijens, J. A., Aabidine, A. Z. El, Rivals, E., & Schönhuth, A. (2017). De novo assembly of viral quasispecies using overlap graphs. *Genome Research*, *27*(5), 835–848. https://doi.org/10.1101/gr.215038.116

Baaijens, J. A., Van der Roest, B., Köster, J., Stougie, L., & Schönhuth, A. (2019). Full-length *de novo* viral quasispecies assembly through variation graph construction. *Bioinformatics*, *35*(24), 5086–5094. https://doi.org/10.1093/bioinformatics/btz443

Charles, O. J., Venturini, C., Goldstein, R. A., & Breuer, J. (2024). HerpesDRG: a comprehensive resource for human herpesvirus antiviral drug resistance genotyping. *BMC Bioinformatics*, *25*(1), 279. https://doi.org/10.1186/s12859-024-05885-5

FluSurver. (n.d.). *FluSurver website*. Http://Flusurver.Bii.a-Star.Edu.Sg.

Fong, W., Rockett, R. J., Agius, J. E., Chandra, S., Johnson-Mckinnon, J., Sim, E., Lam, C., Arnott, A., Gall, M., Draper, J., Maddocks, S., Chen, S., Kok, J., Dwyer, D., O'Sullivan, M., & Sintchenko, V. (2023). SABRes: in silico detection of drug resistance conferring mutations in subpopulations of SARS-CoV-2 genomes. *BMC Infectious Diseases*, *23*(1), 303. https://doi.org/10.1186/s12879-023-08236-6

Fourati, S., Reslan, A., Bourret, J., Casalegno, J.-S., Rahou, Y., Softic, L., Chollet, L., Trémeaux, P., Imbert-Marcille, B.-M., Pillet, S., Veyrenche, N., Boudet, A., Deroche, L., Burrel, S., Mouna PharmD, L., Cocherie, T., Schnuriger, A., Pronier, C., Handala PharmD, L., … Rameix Welti, M.-A. (2026). *Real-world emergence of nirsevimab resistance in RSV-B breakthrough infections on behalf of the POLYRES investigators**. https://ssrn.com/abstract=5427106

Fritz, A., Bremges, A., Deng, Z.-L., Lesker, T. R., Götting, J., Ganzenmueller, T., Sczyrba, A., Dilthey, A., Klawonn, F., & McHardy, A. C. (2021). Haploflow: strain-resolved de novo assembly of viral genomes. *Genome Biology*, *22*(1), 212. https://doi.org/10.1186/s13059-021-02426-8

Garriga, C., & Menéndez-Arias, L. (2006). DR_SEQAN: a PC/Windows-based software to evaluate drug resistance using human immunodeficiency virus type 1 genotypes. *BMC Infectious Diseases*, *6*, 44. https://doi.org/10.1186/1471-2334-6-44

Grubaugh, N. D., Gangavarapu, K., Quick, J., Matteson, N. L., De Jesus, J. G., Main, B. J., Tan, A. L., Paul, L. M., Brackney, D. E., Grewal, S., Gurfield, N., Van Rompay, K. K. A., Isern, S., Michael, S. F., Coffey, L. L., Loman, N. J., & Andersen, K. G. (2019). An amplicon-based sequencing framework for accurately measuring intrahost virus diversity using PrimalSeq and iVar. *Genome Biology*, *20*(1), 8. https://doi.org/10.1186/s13059-018-1618-7

Katoh, K., Rozewicki, J., & Yamada, K. D. (2019). MAFFT online service: multiple sequence alignment, interactive sequence choice and visualization. *Briefings in Bioinformatics*, *20*(4), 1160–1166. https://doi.org/10.1093/bib/bbx108

Lemoine, F., & Gascuel, O. (2021). Gotree/Goalign: toolkit and Go API to facilitate the development of phylogenetic workflows. *NAR Genomics and Bioinformatics*, *3*(3). https://doi.org/10.1093/nargab/lqab075

Li, H. (2021). New strategies to improve minimap2 alignment accuracy. *Bioinformatics*, *37*(23), 4572–4574. https://doi.org/10.1093/bioinformatics/btab705

Salvatierra, K., & Florez, H. (2016). Biomedical Mutation Analysis (BMA): A software tool for analyzing mutations associated with antiviral resistance. *F1000Research*, *5*, 1141. https://doi.org/10.12688/f1000research.8740.2

Tzou, P. L., Tao, K., Pond, S. L. K., & Shafer, R. W. (2022). Coronavirus Resistance Database (CoV-RDB): SARS-CoV-2 susceptibility to monoclonal antibodies, convalescent plasma, and plasma from vaccinated persons. *PloS One*, *17*(3), e0261045. https://doi.org/10.1371/journal.pone.0261045

Virus French Database. (n.d.). *Virus French Database - RSV resistance list*. Https://Virusfrenchresistance.Org/Virus-French-Resistance-Rsv/.

WHO. (n.d.). *WHO - Influenza Antiviral susceptibility*. Https://Www.Who.Int/Teams/Global-Influenza-Programme/Laboratory-Network/Quality-Assurance/Antiviral-Susceptibility-Influenza.

Fast and robust graph construction from KEGG metabolic and genomic data

Florent Cabret[1], Ronan Bocquillon[1], and Emmanuel Néron[1]

https://doi.org/10.5281/zenodo.19036695

Abstract

The Kyoto Encyclopedia of Genes and Genomes (KEGG) enables the modelling of biological systems as integrated graphs, where directed metabolic networks capture reaction flows and undirected genomic graphs encode chromosomal proximity of enzyme coding genes. However, constructing these graphs at the scale of complete bacterial organisms remains challenging due to two limitations: web based API access is constrained by rate limits triggering temporary IP blocks, and existing tools are typically restricted to the scale of pathway analysis. Here we present a framework for the fast and robust construction of whole organism metabolic and genomic graphs from KEGG data. Our approach combines a concurrency model with proxy rotation to overcome web access bottlenecks, specialised parsers for tab-delimited and KGML formats, and an embedded analytical database for efficient storage. For graph construction, we introduce optimised methods that handle real world data complexities, including multiple genomic intervals per gene and uncertain boundary annotations. This enables, to the best of our knowledge, the first systematic construction of integrated graphs at the scale of complete bacterial genomes and entire metabolisms. By overcoming previous scalability barriers, our framework provides the foundation for applying advanced trail finding algorithms to reveal how metabolic function and genomic organisation co-evolve across the bacterial domain, moving beyond isolated pathways towards whole-organism comparative analysis.

Keywords: Systems Biology, Metabolic Networks and Pathways, Genome Bacterial, Data Collection, Graph Theory

[1]Laboratoire d'Informatique Fondamentale et Appliquée de Tours, 64 Avenue Jean Portalis, 37200 Tours, France

Correspondence
ronan.bocquillon@univ-tours.fr,emmanuel.neron@univ-tours.fr

Introduction

The Kyoto Encyclopedia of Genes and Genomes (KEGG) was established in 1995 as a knowledge base for the systematic analysis of gene functions, with the ambitious goal of linking genomic information to higher order functional insights (Ogata et al., 1999). From its inception, KEGG has been structured around three core databases: the PATHWAY database for the graphical representation of biochemical pathways, the GENES database for consistently annotated gene catalogs from complete genomes, and the LIGAND database for chemical compounds and enzymes. By integrating genomic, chemical, and systemic functional information, KEGG serves as an indispensable reference for interpreting sequence data and understanding cellular processes, providing a foundational resource for applications ranging from the reconstruction of metabolic pathways to the analysis of gene expression profiles.

The utility of KEGG extends far beyond simple data retrieval. As demonstrated in studies such as Zaharia et al., 2019, KEGG data enables sophisticated integrative analyses that simultaneously consider metabolic pathways and genomic context. Their CoMetGeNe pipeline leverages KEGG's rich annotations to identify conserved metabolic and genomic patterns across multiple species, revealing how evolutionary pressures shape the organization of enzyme coding genes and their corresponding reactions. Such approaches exemplify the growing recognition that biological function emerges not from isolated components but from the complex interplay between metabolic networks and genomic architecture.

Accessing KEGG data programmatically has been facilitated through various software tools over the years. The KEGG API provides RESTful web services that allow automated retrieval of database entries, forming the backbone for numerous analysis pipelines. Early tools like KEGGgraph (Zhang and Wiemann, 2009) introduced graph based approaches to pathway analysis in R, capturing pathway topology for the first time. Specialized query systems such as IsoKEGG (Sultana et al., 2010, 2014) further expanded the possibilities by enabling logic based pathway queries using subgraph isomorphism, allowing researchers to retrieve pathways based on structural patterns. Integration tools like BiKEGG (Jamialahmadi et al., 2016) have bridged KEGG with other databases such as BiGG, facilitating cross database analyses. More recently, packages such as kegg_pull (Huckvale and Moseley, 2023) have improved accessibility through robust Python implementations that handle multiprocessing and fault tolerant retrieval. However, these tools share fundamental limitations inherent to web-based data access, namely rate limits that can result in temporary IP blocks, combined with a lack of built-in proxy support for large-scale concurrent requests. While kegg_pull's authors provide recommendations for sleep times to mitigate throttling, such measures necessarily trade speed for reliability, and the underlying constraints of operating within KEGG's API limits remain unchanged.

Despite these persistent constraints on data access, a particularly powerful paradigm that has emerged from KEGG's pathway representation is the modeling of biological systems as graphs. In this framework, metabolic pathways are represented as directed graphs where vertices correspond to enzymes (or reactions) and arcs represent substrate-product relationships between sequential reactions. Concurrently, genomic context can be modeled as an undirected graph where edges connect genes that are physically close on the chromosome. This dual graph representation (Zaharia et al., 2019) captures the fundamental insight that enzymes catalyzing successive reactions are often encoded by neighboring genes, reflecting evolutionary optimization of co-expression and functional coordination. The graph abstraction enables rigorous computational approaches: searching for maximum span trails in the metabolic graph whose vertex sets induce a single connected subgraph in the genomic graph reveals functionally coherent modules that transcend traditional operon definitions. These trails can subsequently be used for cross-species comparison, by grouping them based on shared reactions or conserved gene neighborhoods, to reconstruct phylogenetic relationships and trace the evolutionary history of metabolic pathways.

Despite the conceptual elegance of graph based representations, existing implementations have been constrained by computational complexity and scope. Early algorithms such as HNET (Zaharia et al., 2019) demonstrated the feasibility of identifying conserved metabolic genomic patterns but were limited to pathway by pathway analysis. Subsequent work by Ahmed Sidi,

2022; Ahmed Sidi et al., 2025 developed more efficient exact methods based on integer linear programming and constraint programming, demonstrating improved scalability on individual pathway instances. However, the $\mathcal{NP}$-hard complexity of the underlying problems imposed fundamental scalability limitations, preventing exact methods from extending beyond isolated pathway boundaries to encompass entire metabolic networks. Moreover, the reliance on KEGG's predefined pathway boundaries potentially fragments biologically meaningful patterns that span multiple pathways.

Recent algorithmic advances by Cabret et al., 2025 have improved this situation. Their binary integer programming model for finding maximum coverage trails in metabolic networks subject to genomic constraints achieves computational speedups of nearly 3000 times compared to previous approaches, allowing it to scale to networks with thousands of vertices. This development now makes it possible to begin systematic construction and analysis of integrated metabolic genomic graphs at the scale of complete bacterial genomes and full metabolisms. By considering all reactions and genes simultaneously, rather than focusing on individual pathways, this approach may help uncover broader organizational principles and evolutionary patterns that previous computational limitations made difficult to access. Additionally, their approach to generating biologically realistic benchmark instances through gamma-distributed degree sampling provides a way to create validation data that reflects true biological network properties. This offers an alternative to relying solely on simplified random graph models.

The convergence of rich KEGG data resources, sophisticated graph theoretic modeling frameworks, and increasingly more powerful algorithmic solvers creates an unprecedented opportunity to understand how metabolic function and genomic organization co-evolve across the bacterial domain. Realizing this opportunity, however, requires robust infrastructure for constructing integrated metabolic genomic graphs from KEGG data that can leverage these new computational capabilities. In this paper, we present a novel approach that addresses this need through two key contributions:

- Section 2 details a faster and more robust methodology for retrieving KEGG data, which overcomes the bottlenecks of web-based access through a concurrency model with proxy rotation and efficient parsing;
- then Section 3 introduces a graph construction framework operating at the scale of whole metabolisms and complete genomes, introducing optimized methods for building both directed metabolic graphs and undirected genomic graphs that handle real-world data complexities.

By moving beyond previous pathway-centric methods, this combined framework enables the application of advanced trail finding algorithms to complete bacterial systems.

Retrieving data from KEGG

To enable the scalable and robust construction of integrated metabolic genomic graphs at the scale of complete bacteria, our data retrieval pipeline is engineered as a cohesive system of four interdependent components. The process begins by programmatically contacting the KEGG REST API through a dedicated interface, after which the returned data, ranging from tab-delimited records to complex XML based KGML files, is parsed using specialized libraries tailored to each format. To overcome the inherent limitations of web-based access, including rate limits and the risk of temporary IP blocks, a sophisticated concurrency model manages thousands of parallel requests through a lock-free proxy rotation scheme, ensuring both speed and reliability. Finally, all retrieved and parsed data is efficiently stored in an embedded analytical database, where a carefully designed schema and bulk insertion strategy preserve referential integrity while maximizing throughput. Together, these elements form the foundation for acquiring the comprehensive, high quality datasets required for downstream graph analysis.

Concurrency model

The large number of independent web requests required to retrieve complete KEGG data for thousands of organisms makes concurrency essential. We therefore implement a thread pool to

manage these requests. Since most of the time in each request is spent waiting for a response from the KEGG server, the pool is deliberately oversized, and we create 25 times the number of logical processor threads. This over-subscription ensures that while some threads are blocked on I/O, others can continue making progress, thereby maximising network utilisation.

A critical challenge when issuing many requests to the same public API is the risk of temporary IP blocks. To mitigate this, we obtain before data retrieval begins a list of public proxies drawn from monosans/proxy-list, an hourly updated repository of freely available proxies scraped from diverse sources and validated for reliability. Each proxy is then tested for reachability, requiring a successful connection attempt to the KEGG website within 10 seconds and allowing up to three retries. Proxies that pass this filter are stored in a global list alongside a flag indicating whether they are currently in use and an optional timestamp of the last failure. This list is fixed for the duration of the run and provides the pool of proxies used for all requests.

Access to this shared list must be efficient and must not become a bottleneck. We therefore implement a lock-free proxy rotation scheme. Lock-free algorithms guarantee system-wide progress even if individual threads are delayed, and they avoid the overhead of kernel-space locking. Specifically we use the test and test-and-set idiom. A thread repeatedly reads the state of a randomly chosen proxy; only if the proxy appears free does it attempt to claim it with an atomic exchange. If the exchange succeeds, the thread proceeds to use the proxy. Should the request later fail with an HTTP status of 400 or higher, which indicates a client or server error, the proxy is marked as temporarily unavailable and a 5 second timeout is recorded. This timeout value follows the recommendation of kegg_pull to avoid hammering a faulty proxy. Because the entire selection loop uses only relaxed-order loads and a single acquire-release exchange, contention on the proxy list is negligible.

Each request itself follows a robust retry policy. After a proxy has been acquired, the thread issues a GET request with a generous one minute timeout. If the request times out or returns an error status, it is retried up to three times. A failed request does not immediately discard the proxy; instead the proxy is temporarily penalised (the 5 second timeout) so that other proxies can be tried, while a faulty proxy is given time to recover.

Retrieved responses must be stored in a relational database while preserving referential integrity. Because foreign keys must point to existing primary keys, we enforce a strict ordering of insertions using the bulk synchronous parallel (BSP) model. The workflow is divided into super-steps: first an organism is inserted, then all genes and positions, then enzyme annotations, and finally the metabolic graph edges (see Figure 1). A barrier synchronises all worker threads at the end of each super-step. The barrier itself is implemented with the low-level atomic_wait and notify_all primitives. Threads that finish early start the next super-step up to the point of committing to the database; if the remaining threads have not yet reached the barrier, they first spin briefly before finally parking on the atomic flag. This hybrid approach avoids excessive context switching when the last thread is expected to arrive quickly, while still allowing long waits to block efficiently. This design guarantees that no thread ever tries to insert a row that references a still missing primary key.

To further reduce contention, each thread maintains its own database connection in thread-local storage. Instead of issuing individual INSERT statements (which would require repeated round trips and serialisation), threads accumulate rows in memory and commit them in bulk transactions. This approach minimises lock contention on the database and significantly cuts the number of inter-thread synchronisation events. Only when a barrier is reached must the threads ensure that all their buffered writes have been applied; the atomic wait barrier provides exactly that guarantee.

Contacting the API

To programmatically access the KEGG REST API, we are using libcurl, a free and portable client-side URL transfer library. It supports a wide range of protocols including HTTPS, which is required for all of the endpoints, and provides robust features such as persistent connections, thread safety, and comprehensive error handling. Its ability to seamlessly handle HTTP and SOCKS proxies, including those requiring authentication, is particularly valuable in institutional

and high-performance computing environments where direct external access may be restricted. By leveraging libcurl, our workflow can reliably retrieve genomic, pathway, and chemical data from KEGG while automatically routing traffic through necessary proxy servers, ensuring both reproducibility and resilience in diverse network settings.

Parsing CSV and XML

The KEGG API returns data in several distinct formats depending on the operation performed. Most operations including list, find, conv, and link return tab-delimited text that can be parsed line by line, while pathway data retrieved via the get operation with the kgml option is provided in the XML-based KGML (KEGG Markup Language) format. These two structural paradigms require different parsing strategies: the former demands efficient, type-safe handling of flat records, while the latter requires navigating hierarchical relationships between pathway elements. To address these requirements, we employ two specialized C++ libraries. For tab-delimited responses, we leverage the PEG based parser combinator library lexy to construct declarative grammars that map directly onto data structures. For KGML, we utilize the DOM interface provided by pugixml to traverse and extract pathway components.

Parsing Expression Grammars (PEGs) provide a formal and unambiguous method for describing machine oriented syntax using ordered choice operators (Ford, 2004). This paradigm was notably adopted by the CPython interpreter in Python 3.9 (PEP 617) to lift the LL(1) restrictions and simplify grammar maintenance. The practical advantages of this approach are now being demonstrated in production database systems, with the recently released DuckDB v1.5.0 adopting a PEG based parser to enable runtime extensibility and deliver significantly improved error messages and auto-completion. This validation in a high-performance setting reinforces our choice of a PEG based solution for the KEGG parsing task. To process KEGG's tab-delimited outputs, we implement this approach using lexy, a C++ parser combinator library. Its expressive domain-specific language embeds grammars within the code, mirroring the structure of handwritten recursive descent parsers while providing explicit control over backtracking through branch conditions. This design avoids the performance pitfalls common in traditional PEG implementations. By constructing specialized grammars, we ensure type-safe and zero-copy parsing, complete with robust error recovery to handle the irregular formatting sometimes found in biological data dumps.

To process KGML outputs, we employ pugixml, a lightweight and high performance XML parsing library. This format encodes pathway information in a hierarchical structure that defines elements such as genes, compounds, reactions, and their relations. The library's DOM interface provides intuitive traversal methods, enabling efficient navigation of the KGML structure and ensuring reliable data extraction for subsequent analysis through the retrieval of pathway components with minimal overhead.

Database management system

DuckDB is a fast analytical, in-process, open-source and portable database system. Following the paradigm popularized by SQLite, it operates embedded within a host application, eliminating external dependencies and the overhead of inter-process communication. Its architecture is specifically optimized for Online Analytical Processing (OLAP) workloads through a columnar-vectorized query execution engine, which processes data in large batches rather than row-by-row to accelerate complex aggregations and joins on massive datasets. This combination of embedded simplicity and analytical performance makes DuckDB particularly well suited for efficient data analysis within the application itself. To support this analysis, the database is structured according to the entity-relationship model illustrated in Figure 1. Such a unified storage could also benefit tools like KEGG spider (Antonov et al., 2008), which builds a global gene metabolic network by integrating all the pathways, enabling the interpretation of gene lists in a pathway-free context. Rather than treating each pathway as an independent unit, it constructs a unified network where genes are connected if their associated reactions share common compounds. It then infers network models from an input gene list by progressively connecting genes at increasing distances and assesses the statistical significance of these models through Monte Carlo simulations. This

approach offers an alternative to the directed metabolic graph construction presented later in this paper.

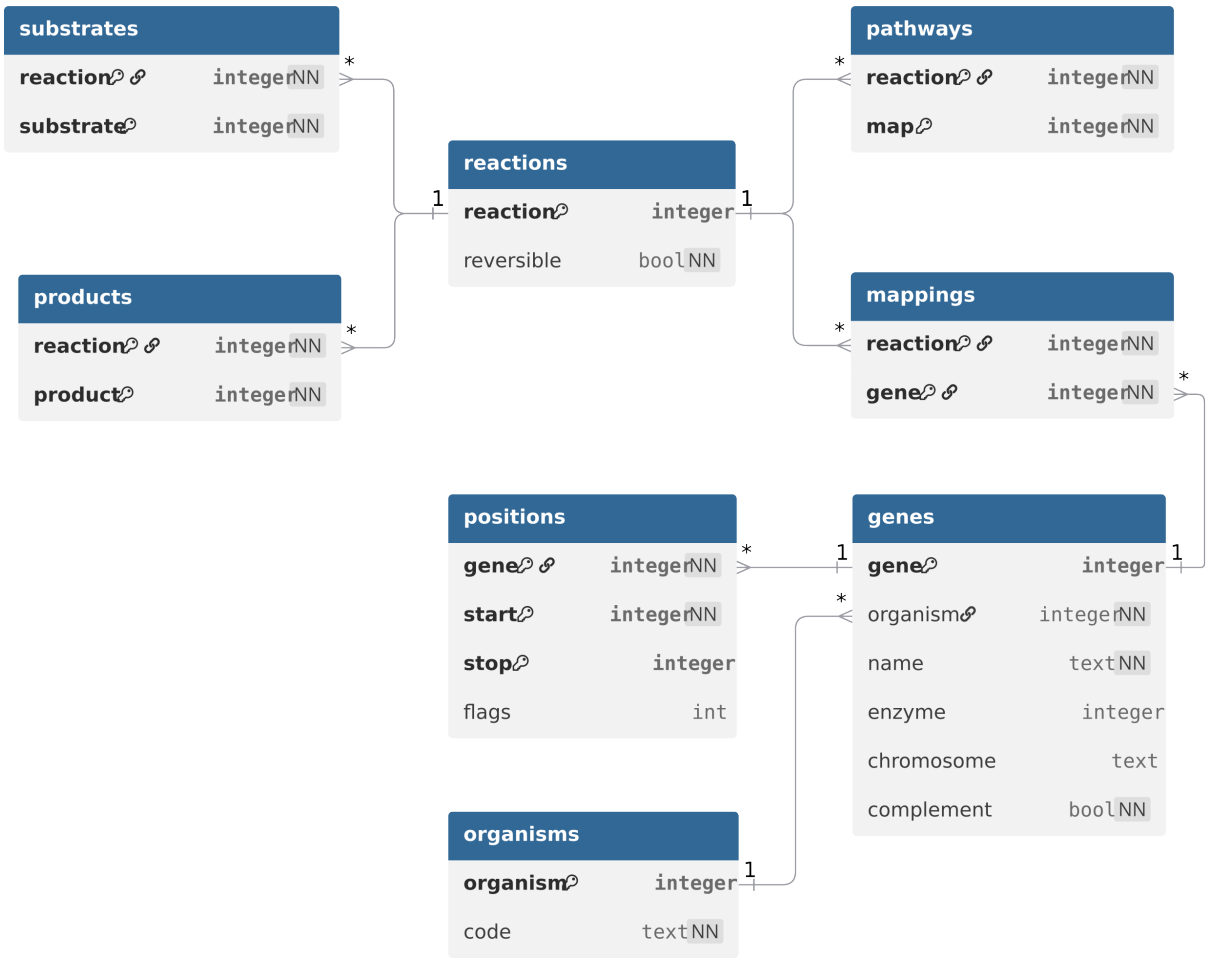


Figure 1 – Entity-relationship diagram of the database

Numerical experiments

Retrieving complete KEGG data for 1000 bacterial organisms with our approach requires approximately one hour of total runtime, with a variance of about ± 15 minutes attributable to the heterogeneous performance of public proxies (since all threads must synchronise at each super-step, a slower proxy can delay every worker). In practice, the number of proxies can be reduced to avoid placing an excessive load on the KEGG server. For smaller analyses involving at most 100 organisms, the parallel portion of the workload is substantially reduced. Amdahl's law (Amdahl, 1967) then predicts that the fixed overheads, notably the two minutes needed for initial proxy filtering, dominate the total time, making the concurrent strategy less beneficial than a simpler sequential retrieval.

Regarding long term robustness against changes in KEGG's access policies, our pipeline is designed to accommodate progressive hardening of the API. Rate limits and temporary IP blocks are already handled by proxy rotation and retries. Should KEGG introduce more advanced challenges such as CAPTCHAs or JavaScript based access gates, these can be integrated by adding a headless browser automation layer coupled with a CAPTCHA solving service. While such extensions would increase latency, they would preserve the overall architecture and concurrency model. For non interactive bulk access, KEGG also provides a paid FTP subscription that delivers flat files; our parser components (PEG for tabular data, pugixml for KGML) can ingest those offline files unchanged, bypassing API restrictions entirely. Thus, the framework is not locked into a

single acquisition strategy but can be adapted to alternative data sources or evolving server policies through interchangeable backend modules.

In terms of data quality, our pipeline issues more API calls than the method of Zaharia et al., 2019 because it fetches complete, fine-grained data for every organism instead of relying on global association tables. The additional calls are the price paid for obtaining precise reaction-enzyme associations and handling complex genomic annotations, which the earlier approach approximated or omitted.

Construction of the graphs

The core of the analysis framework lies in constructing two complementary graph representations from the retrieved KEGG data: a directed metabolic graph capturing the flow of substrates and products between reactions, and an undirected genomic graph encoding the chromosomal proximity of the associated enzyme-coding genes. These constructions follow the foundational concepts introduced by Zaharia et al., 2019, which enable the integrated analysis of metabolic and genomic organization. To achieve the necessary performance for organism scale analysis, we have developed optimized variants of these foundational constructions. The following subsections describe each graph construction in detail, using side by side comparison with the original methods to highlight how our approach overcomes previous scalability limitations.

Metabolic graph

The metabolic graph is a directed graph that models the flow of metabolites through a sequence of reactions. In Zaharia et al., 2019, its construction begins by representing both reactions and chemical compounds as nodes in a bipartite graph. Directed edges are placed from substrate compounds to the reactions they participate in, and from each reaction to its product compounds. The final directed graph is obtained by projecting this bipartite structure onto the set of reaction nodes. In this projection, two reactions become connected by a directed arc if a product of the first reaction serves as a substrate for the second. This results in a pathway centric directed graph where arcs trace feasible metabolic sequences, abstracting away the underlying compounds. KEGG provides a reversible flag as a boolean attribute for each reaction. For reactions marked as reversible, the projection naturally creates arcs in both directions because each compound is considered both a substrate and a product. Thus, two reactions linked via a reversible reaction may have arcs in both directions after projection.

In this work, we adopt a more efficient construction that bypasses the explicit bipartite graph and its projection. Instead, we maintain two hash maps: one that associates each compound with the set of reactions that produce it, and another that associates each compound with the set of reactions that consume it. For every compound, we then iterate over all producer-consumer pairs and directly add a directed arc from the producer to the consumer whenever the product of the former serves as a substrate of the latter. This approach eliminates the storage of intermediate bipartite edges and the overhead of a separate projection step. During arc creation, we also query the enzymes associated with each reaction and discard any reaction that lacks an enzyme annotation, ensuring that only reactions with genomic support are retained in the final graph. The resulting directed graph retains the same biochemical interpretation, arcs represent feasible metabolic transitions, but is constructed in a single pass with reduced memory footprint and lower time complexity, making it particularly suitable for large-scale metabolic network analysis.

Genomic graph

The genomic graph is an undirected graph that captures the spatial proximity of genes encoding the enzymes of the corresponding reactions. Zaharia et al., 2019 first build this graph by linking consecutive genes on each chromosome, forming a linear chain. For circular bacterial chromosomes, an additional edge connects the first and last gene to correctly reflect the circular topology. By default, the graph is constructed separately for genes on the forward and reverse strands, reflecting the biological observation that functionally related genes tend to reside on the same strand. This behavior can be modified to consider genes on both strands together when

a more permissive neighborhood definition is desired. To allow for gaps between functionally related genes, the shortest path distances in this initial gene graph are computed. If two genes are separated by at most a specified number δ of intermediate genes along the chromosome, an edge is added directly between them (typically, $\delta \leq 3$). Reactions are then mapped to the genes that encode their enzymes. Two reactions are linked by an undirected edge precisely when their associated genes are neighbors in the graph. This genomic graph thus encodes the condition that the genes of a metabolic trail must lie in a genomically coherent cluster.

In this work, we employ an Order Statistic Tree using as the base data structure the Logarithmic Binary Search Tree introduced by Roura, 2001. In this structure, the size of the subtree stored at each node serves the dual purpose of both maintaining the balancing property and enabling efficient rank queries. Each gene can be associated with multiple genomic intervals due to the presence of introns or uncertainties arising from high-throughput sequencing errors. In our database schema (see Figure 1), a gene can thus be associated with several position entries. Other complications include entries where the stop value is missing, in which case we set it equal to the start value. Additionally, a flag field, which we currently ignore, might indicate that the true start or stop lies outside the reported values.

The positions of all genes are inserted into the tree ordered by their start value. The data structure provides logarithmic time rank queries, which we use to compute genomic distances efficiently. For a given pair of genes, we first compute a new interval spanning from the smallest start value to the largest stop value among all intervals belonging to that gene. If these new intervals don't intersect, the genomic distance between the two genes is simply the distance between them. This value can be obtained directly using rank queries on the tree without examining individual intervals.

When they overlap, a more detailed analysis is required. We apply a sweep line algorithm on the sorted list of all interval endpoints belonging to the two genes. The algorithm maintains two pointers, each initially pointing to the first interval of the respective gene. At each step, it compares the current intervals: if they intersect, the distance is zero; otherwise, it advances the pointer corresponding to the interval with the smaller start value. This process continues until either an intersection is found or all intervals have been processed, yielding the minimal distance between any two positions belonging to the two genes. By leveraging the Order Statistic Tree for initial sorting and span checks, and the sweep line for overlapping cases, we achieve an efficient and accurate computation of gene neighbourhood distances even in the presence of multiple position intervals per gene.

It is worth noting that the genes encoding enzymes represent only a small subset of all coding genes in a genome. This specificity makes the construction of the genomic graph more computationally manageable and focuses the analysis on biologically relevant functional units. Compared to the approach of Zaharia et al., 2019, which does not account for the complexities introduced by multiple intervals per gene, missing stop values, or flag attributes indicating uncertain boundaries, our method is designed to handle these real-world data imperfections robustly. By efficiently computing distances between genes even in the presence of such complications, our approach provides a reliable foundation for downstream analyses. This capability could prove particularly useful for trail comparison if the need for less strict genomic constraints were to arise, thereby supporting scalable whole genome analyses.

To go further, the genomic graph, which captures chromosomal proximity between enzyme-coding genes, could be refined by integrating spatial proximity data derived from chromosome conformation capture techniques such as Hi-C. The contact matrix used by Gao et al., 2024 provides genome-wide interaction frequencies between genomic loci, revealing that genes physically close in the three-dimensional space of the nucleoid are more likely to be functionally associated, even when they are far apart on the linear chromosome. By incorporating these spatial interaction frequencies as additional edges in the genomic graph, the resulting network could better reflect the functional organisation of the genome, complementing our method. Moreover, manual annotations remain valuable for correcting annotation errors, resolving ambiguous gene boundaries, or incorporating experimentally validated interactions that may not be captured by high-throughput data.

Numerical experiments

To illustrate the scalability of our graph construction pipeline (see Table 1), we built metabolic and genomic graphs for six bacterial species with diverse number of reactions: the reduced genome of the secondary endosymbiont of *Trabutina mannipara* (senm), the *Blochmannia* endosymbiont of *Camponotus (Colobopsis) obliquus 757* (ben), the porcine pathogen *Glaesserella parasuis SH03* (hpas), the marine bacteria *Vibrio campbellii ATCC BAA-1116* (vac), the model organism *Escherichia coli K-12 MG1655* (eco), and the opportunistic pathogen *Klebsiella grimontii* (kgr). Construction of both graphs from the retrieved KEGG data scales with the organism complexity and consistently takes under 30 seconds. While the six species studied here span a wide range of reaction counts, the average across all bacteria is 946. The experiments confirm that finding a maximum span trail in these graphs (*i.e.*, at the scale of a complete bacterial organism) is tractable, even if it requires more time than expected. In contrast, Cabret et al., 2025 solve this same problem on a set of biologically realistic synthetic graphs of 1060 vertices in less than 3 seconds on average, and they obtain longer trails (36 reactions on average). These discrepancies highlight the computational challenge of the underlying combinatorial problem, as well as the difficulties to create an accurate model of synthetic graphs.

Bacteria name		senm	ben	hpas	vca	eco	kgr
Number of reactions		26	340	647	958	1130	1269
Metabolic graph	order	17	309	594	863	1019	1161
	size	22	538	1484	2737	3578	4071
Genomic graph	size	26	1377	2091	2489	2620	2899
Construction time (s)		0.3	4	9	15	22	26
Maximum span trail		5	10	25	18	12	12
Resolution time (s)		0.05	2	7	29	83	89

Table 1 – Statistics for different bacterial species

Conclusion

We have presented a framework for constructing integrated metabolic and genomic graphs from KEGG data at the scale of complete bacterial organisms. Our approach addresses the bottlenecks of web-based data retrieval through a concurrency model with proxy rotation. Our graph construction pipeline leverages two specialized data structures: hash maps for efficient metabolic graph assembly and order statistic trees for genomic proximity computations. To the best of our knowledge, this enables the first systematic construction of whole organism graphs that were previously limited to pathway by pathway analysis.

With these organism scale graphs and the trails they support, the next challenge becomes comparative analysis across species. Zaharia et al., 2019 provided an elegant method for identifying conserved patterns. However, their approach reintroduces a segregation between reactions and genes during comparison: trails are treated as reaction sets, temporarily decoupling the metabolic and genomic dimensions that the model itself integrates.

Although our framework is presented using KEGG, the modular design of its components (the concurrency model with proxy rotation, the PEG based tabular parser, the XML parser, and the embedded analytical database) makes them readily adaptable to other biological databases that expose REST APIs or provide structured flat files. For instance, replacing the KEGG API endpoints with those of any alternative data source would require only a thin configuration layer, namely URL templates and response grammar definitions. Similarly, the graph construction algorithms operate on generic reaction compound and gene interval data models, and can therefore be applied to any metabolic and genomic dataset that adheres to the same relational schema. This generality is explicitly supported by the use of DuckDB, which allows the user to load alternative data sources into the same table structure.

The availability of whole organism graphs now calls for comparison methods that preserve this duality. We need approaches that identify conserved substructures simultaneously in both metabolic and genomic dimensions. Such methods would reveal not only which reaction sequences persist across evolution, but also how their underlying genomic organization is maintained or rearranged. Developing these methods represents the next frontier in understanding the co-evolution of metabolism and genome architecture across the bacterial domain.

Fundings

The authors declare that they have received no specific funding for this study.

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.

Data, script, code, and supplementary information availability

Data, script and codes are available online (https://doi.org/10.5281/zenodo.19025626; Cabret et al., 2026).

References

Ahmed Sidi ML (2022). Nouvelles approches informatiques et mathématiques pour la résolution de problèmes biologiques. PhD thesis. Université de Tours - LIFAT. URL: https://hal.science/tel-03807522.

Ahmed Sidi ML, Bocquillon R, Cabret F, Mohamed Babou H, Dhib C, Néron E, Soukhal A, Nanne MF (2025). Constraint Programming Approaches for Finding Conserved Metabolic and Genomic Patterns. *Computers & Operations Research* **183**, 107166. https://doi.org/10.1016/j.cor.2025.107166.

Amdahl GM (1967). Validity of the Single Processor Approach to Achieving Large Scale Computing Capabilities. In: *Proceedings of the April 18-20, 1967, Spring Joint Computer Conference*. AFIPS '67 (Spring). New York, NY, USA: Association for Computing Machinery, pp. 483–485. https://doi.org/10.1145/1465482.1465560.

Antonov AV, Dietmann S, Mewes HW (2008). KEGG Spider: Interpretation of Genomics Data in the Context of the Global Gene Metabolic Network. *Genome Biology* **9**, R179. https://doi.org/10.1186/gb-2008-9-12-r179.

Cabret F, Bocquillon R, Neron E (2025). Towards Mining Neighbourhood Patterns by Crossing the Metabolic and Genomic Networks at the Organism Level. Unpublished manuscript.

Cabret F, Bocquillon R, Neron E (2026). *KEGG Extractor: A Toolkit for Data Retrieval and Construction of Whole-Organism Metabolic and Genomic Graphs*. Zenodo. https://doi.org/10.5281/zenodo.19025627.

Ford B (2004). Parsing Expression Grammars: A Recognition-Based Syntactic Foundation. In: *Proceedings of the 31st ACM SIGPLAN-SIGACT Symposium on Principles of Programming Languages*. POPL '04. New York, NY, USA: Association for Computing Machinery, pp. 111–122. https://doi.org/10.1145/964001.964011.

Gao Y, Ma B, Xu Q, Peng Y, Gong H, Guan A, Hua K, Langford PR, Jin H, Luo R (2024). Spatial Proximity and Gene Function: A New Dimension in Prokaryotic Gene Association Network Analysis with 3D-GeneNet. *Briefings in Bioinformatics* **25**, bbae320. https://doi.org/10.1093/bib/bbae320.

Huckvale E, Moseley HNB (2023). Kegg_pull: A Software Package for the RESTful Access and Pulling from the Kyoto Encyclopedia of Gene and Genomes. *BMC Bioinformatics* **24**, 78. https://doi.org/10.1186/s12859-023-05208-0.

Jamialahmadi O, Motamedian E, Hashemi-Najafabadi S (2016). BiKEGG: A COBRA Toolbox Extension for Bridging the BiGG and KEGG Databases†. *Molecular BioSystems(MBS)* **12**, 3459–3466. https://doi.org/10.1039/C6MB00532B.

Ogata H, Goto S, Sato K, Fujibuchi W, Bono H, Kanehisa M (1999). KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acids Research* **27**, 29–34. https://doi.org/10.1093/nar/27.1.29.

Roura S (2001). A New Method for Balancing Binary Search Trees. In: *Automata, Languages and Programming*. Ed. by Fernando Orejas, Paul G. Spirakis, and Jan van Leeuwen. Berlin, Heidelberg: Springer, pp. 469–480. https://doi.org/10.1007/3-540-48224-5_39.

Sultana KZ, Bhattacharjee A, Jamil H (2010). IsoKEGG: A Logic Based System for Querying Biological Pathways in KEGG. In: *2010 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*. 2010 IEEE International Conference on Bioinformatics and Biomedicine (BIBM), pp. 626–631. https://doi.org/10.1109/BIBM.2010.5706642.

Sultana KZ, Bhattacharjee A, Jamil H (2014). Querying KEGG Pathways in Logic. *International Journal of Data Mining and Bioinformatics* **9**, 1–21. https://doi.org/10.1504/IJDMB.2014.057772.

Zaharia A, Labedan B, Froidevaux C, Denise A (2019). CoMetGeNe: Mining Conserved Neighborhood Patterns in Metabolic and Genomic Contexts. *BMC Bioinformatics* **20**, 19. https://doi.org/10.1186/s12859-018-2542-2.

Zhang JD, Wiemann S (2009). KEGGgraph: A Graph Approach to KEGG PATHWAY in R and Bioconductor. *Bioinformatics* **25**, 1470–1471. https://doi.org/10.1093/bioinformatics/btp167.

Holograph: a generic RDF schema to handle data from agroecological holobionts

Marie Lahaye[1], Alice Mataigne[2], Edmond Berne[3,4,5], Matéo Boudet[1,2], Maria Bernard[6,7], Christophe Mougel[1], Lionel Lebreton[1], Valentin Loux[3,4], Anne-Françoise Adam-Blondon[5,8], Olivier Rué[3,4] and Fabrice Legeai[1,2]

[1] UMR 1349 IGEPP, INRAE, Le Rheu 35650, France
[2] Univ Rennes, Inria, CNRS, IRISA - UMR 6074, F-35000 Rennes, France
[3] Université Paris-Saclay, INRAE, MaIAGE, 78350 Jouy-en-Josas, France
[4] Université Paris-Saclay, INRAE, BioinfOmics, MIGALE bioinformatics facility, 78350 Jouy-en-Josas, France
[5] IFB-Core, Institut Français de Bioinformatique (IFB), CNRS, INSERM, INRAE, CEA, 94800 Villejuif, France
[6] INRAE, SIGENAE, Jouy-en-Josas, France
[7] INRAE, AgroParisTech, GABI, Université Paris-Saclay, Jouy-en-Josas, France
[8] URGI, INRAE, Université Paris-Saclay, 78026 Versailles, France

Corresponding author: marie.lahaye@inrae.fr, fabrice.legeai@inrae.fr

Keywords data integration, ontologies, holobionts, RDF, agroecology, metagenomics

Abstract *Understanding host-microbiota interactions in agroecological studies requires linking complex heterogeneous datasets, such as microbial diversity, host phenotype and environmental context. In response, we developed Holograph, a RDF (Resource Description Framework) schema dedicated to the representation of holobiont data, to provide a structured and interoperable way to store such heterogeneous information. This schema is structured around three main ontologies: SOSA, I-ADOPT and GeoSPARQL and integrates specific terms from other convenient ontologies. This schema was successfully tested and implemented on datasets describing plant and livestock animal holobionts allowing data exploration through SPARQL queries or web applications.*

Introduction

Holobionts aim to represent a host (an animal or a plant), and its microbiota as a whole. Agroecological projects studying holobionts entail identifying and characterizing the communities of living organisms associated with cultivated plants or livestock. The composition of the microbiota is generally obtained using metabarcoding or metagenomics methods, resulting in abundance tables of Amplicon Sequence Variants (ASVs) or metagenome-assembled genomes (MAGs), representing the observed taxa. In addition, in holobiont studies, many variables are collected in the field or in laboratory to accurately describe the host phenotype (including health status, growth, physiological or performance traits) or host products. Furthermore, the environment can be modelled with descriptors (e.g. presence of bioagressors, agricultural practices including diet composition, climatic data).

To properly manage and capitalize on the complexity of holobionts data, a modelization as a knowledge graph using generic and domain specific ontologies is essential. This representation enables precise semantic description of data and metadata and links information describing the host, the microbiota and the environment. Reusing terms from shared ontologies will allow interoperability between datasets, and other knowledge graphs.

In the frame of the french program MUDIS4LS (Mutualized Digital Spaces for FAIR data in Life and Health Science), we developped Holograph, a generic RDF schema based on the SOSA and I-ADOPT core-ontologies. This schema was applied to integrate data from large-scale projects: DeepImpact and ChickStress/Feed-a-Gene.

Material and methods

RDF

The Resource Description Framework (RDF) is a standard model to describe data as an oriented graph. It is divided in three elementary entities: the subject and the object linked by a predicate, all identified by a Uniform Resource Identifier (URI). This is an interoperable model thanks to URIs and the use of controlled vocabulary and ontologies. An ontology is a description and organisation of concepts to structure knowledge about topics so that actors of a same domain can share the same vocabulary. RDF data are loaded into triplestores such as Virtuoso which allows data querying through the SPARQL query language.

Ontologies

The schema partly relies on a combination of three core ontologies. The **SOSA** core ontology [1] defines classes and properties to describe observations made by sensors on a feature of interest, the observed property and temporal information. The **I-ADOPT** framework ontology[2] is designed to help with interoperability between existing variables by representing them as more specific entities that can match terms described in ontologies or taxonomies. The **GeoSPARQL**[3] ontology defines a vocabulary for representing geospatial data. We only used its two main classes (Feature and SpatialObject) as they were sufficient to describe nested locations. Our schema is also linked to the **NCBITAXON** ontology[4], which represents the NCBI organismal taxonomy. To form the rest of the schema, we extended these core ontologies using specific terms from other relevant ontologies, when available, to describe classes, properties or relations (see Table 1).

Ontology	ID	Description
Plant Ontology[5]	PO	Plant anatomy, morphology, growth and development
Uber-anatomy ontology[6]	UBERON	Cross-species anatomy ontology
Environmental ontology[7]	ENVO	Environments, ecosystems, habitats and related entities
Agronomy Ontology[8]	AGRO	Agronomic pratices, techniques, variables and related entites
FoodOn Food Ontology[9]	FOODON	Farm-to-fork ontology describing food
Relation Ontology[10]	RO	Relationship types destined to be used in multiple ontologies
Thesaurus INRAE[11]		INRAE thesaurus used to index and annotate web pages, datasets or specific terms

Tab. 1. List of the ontologies terms where extracted from to build the schema

Datasets

The model was first tested using the **DeepImpact** dataset as a plant holobiont use case. This project aims to understand the impact of soil's microbial diversity, as well as physico-chemical and environmental conditions on performance of wheat and rapeseed crops. This dataset includes hundreds of agroenvironmental descriptors (e.g. bioagressors inventories, plant phenotypic traits, agricultural practices, soil biochemistry) measured at various spatial levels (field, plot or plant) collected in 4 plots from 200 agricultural fields during 2 years (4 seasons of sampling). This dataset has been associated with climatic data extracted from the MeteoFrance SAFRAN database. In each plot, microbial commu-

nities from plant compartments (leaf, root and rhizosphere) have been sequenced (39 runs covering 6379 samples) and ASV, their taxonomic affiliations and abundances were derived from bioinformatics analyses.

The model was then applied to the dataset from the **ChickStress** and **Feed-a-Gene** projects to generalize it to both plant and livestock holobionts. Part of these projects aim to study the relationship between feed efficiency and gut microbiota in laying chickens under contrasting feeding conditions[12]. This dataset includes measurements on 57 hens on which caecum microbiotas were analysed. The hens belong to two distinct lines (R+ and R-), which are highly divergent in feed efficiency, and were fed either a control diet (CTR) or a fiber-enriched, energy-depleted diet (LE). During and at the end of the experiment, various performance (linked to feed efficiency or egg productions and quality), growth or physiological measures (including blood analysis) were recorded.

Results

Holograph schema

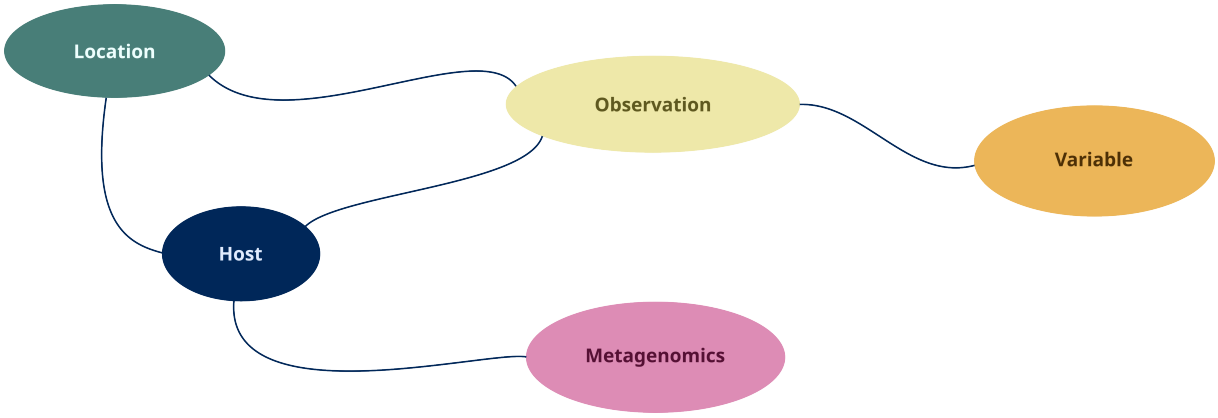


Fig. 1. Overview of the holograph schema

The holograph schema can be broken down into 5 different subparts (see Figure 1). The first one describes the **host**, either an animal or a plant, its products (eggs, in our use case), diet composition and the sampled compartments. The compartment from the chicken dataset was the caecum, term found in the UBERON ontology. In the DeepImpact project the compartments were plant organs: roots, leaves (terms found in the PO) and rhizosphere (ENVO).

These compartments can be sampled for **metagenomics** studies. As this part of the schema did not match any known ontology we developed a new model. It describes the sequenced sample, the run and the analysis, as well as the experiment and the project (if needed). Specific metadata are stored (eg: total read count, sequencing instrument or sequenced marker). Each analysis has results stored under the classes TaxonomicAssignation and Abundance, both linked to an ASV. The taxonomy has a direct link with NCBITAXON as, when instantiating the model, we extract the most accurate taxon URI found in NCBITAXON.

A host is located on a **location** which can be contained in another location. The links between the different locations are represented using the GeoSPARQL ontology. For example, a plant can be located in a plot contained in fields, and a hen is reared in a cage housed in farm. The different types of locations are terms coming from ENVO, AGRO and the Thesaurus INRAE.

Observations such as the host, the diet, the products or the locations are represented using the SOSA ontology. An observation is made by a sensor (a person or a device) on a object. A measured variable, named ObservableProperty in SOSA, is stored as a value and a unit. The observation also has a result time and is linked to a time interval during which the measurement was made (eg: the sampling season or year).

Variables are described as an ObservableProperty in the SOSA ontology. However, modelization can be quite complex (e.g. class density of a weed species at a specific phenological stage). To describe them, and allow interoperability with ontologies, we used the I-ADOPT ontology. Variables are decomposed into elementary entities: a property, an entity and a constraint on this entity. Variables can also be grouped in sets.

Data integration and exploration

To query the model, data must either be loaded directly into a triplestore (eg: Virtuoso) or through a dedicated interface, such as AskOmics (https://askomics.org/). To ease the formatting of user data into the Holograph model, we provided multiple CSV templates and adhoc scripts for conversion into the turtle standard format. Users can directly explore the ingested data with SPARQL queries into the triplestore, or through the AskOmics web interface.

Instantiating the schema with the DeepImpact dataset resulted in 20 210 804 triples and with the ChickStress/Feed-a-Gene dataset 176 190 triples. The datasets can be explored through questions, such as (for DeepImpact): *What are the taxa found in root microbiota of fields in western France under conventional agriculture in the 2nd sampling of the 2nd year*. This resulted in a SPARQL query that returned 1005 results in 4.717 seconds.

Discussion

The current Holograph schema is flexible, as terms can be added to represent other locations, products, sampling compartments. It also can be adapted to other applications: bioreactors, ecosystems or fermented foods. However, any adaptation will be met with some difficulties.

Identifying the best match for terms defined in the schema was one of the main challenge, as terms might be defined across multiple ontologies, and the most suitable descriptions may come from ontologies outside the expected domain. We tried to favor ontologies from our domain or at least generic ones, for more robustness. In particular, in the metagenomics part of the model we did not find terms in convenient ontologies or sometimes the definition of terms were not acceptable (e.g. the term sequencing run is well described in the National Cancer Institute Thesaurus OBO Edition, but this thesaurus is not relevant to agroecological study). It would be appropriate to complement existing ontologies or the thesaurus INRAE with the missing terms and their precise definitions. Consequently, filling the templates requires time to find matching terms to describe variables from the dataset. Yet, it is unavoidable to ensure data and metadata integrity. Thus, a good documentation is essential, as well as some training and support of the data producer.

Querying the model can be challenging: if the data are loaded into an or Virtuoso instance, the user needs to know how to write SPARQL queries. If using an AskOmics instance, SPARQL queries writing can be avoided but the user still needs to know and understand the schema. To answer these

problems, a new user-friendly web application dedicated to the Holograph schema is needed to allow the user to query the model easily, without having to be familiar with it as the complexity would be hidden. Contrary to AskOmics, this web interface would be specific to the schema and could be extended to other schemas using the SOSA and I-ADOPT ontologies.

Acknowledgments

This work is funded by the French National Research Agency (ANR): France 2030, Investments for the Future Program (ANR-11-INBS-0013), EQUIPEX+ (ANR-21-ESRE-0048), Alternative Crop Production and Protection (ANR-20-PCPA-0004), and BIOADAPT (ANR-13-ADAP). It has also been supported by the European Union's H2020 programme Feed-a-Gene project (No. 633531).

References

[1] Janowicz K, Haller A, Cox SJD, Le Phuoc D, Lefrançois M. SOSA: A lightweight ontology for sensors, observations, samples, and actuators. Journal of Web Semantics. 2019;56:1-10. Available from: https://www.sciencedirect.com/science/article/pii/S1570826818300295.

[2] Magagna B, Moncoiffé G, Devaraju A, Stoica M, Schindler S, Pamment A, et al.. InteroperAble Descriptions of Observable Property Terminologies (I-ADOPT) WG Outputs and Recommendations. Zenodo; 2022. Available from: https://doi.org/10.15497/RDA00071.

[3] Nicholas J Car, Timo Homburg, Matthew Perry, John Herring, Frans Knibbe, Simon J D Cox, et al. OGC GeoSPARQL - A Geographic Query Language for RDF Data. Open Geospatial Consortium; 2023. OGC 22-047. Available from: http://www.opengis.net/doc/IS/geosparql/1.1.

[4] Schoch CL, Ciufo S, Domrachev M, Hotton CL, Kannan S, Khovanskaya R, et al. NCBI Taxonomy: a comprehensive update on curation, resources and tools. Database (Oxford). 2020 Jan;2020.

[5] Cooper L, Walls RL, Elser J, Gandolfo MA, Stevenson DW, Smith B, et al. The plant ontology as a tool for comparative plant anatomy and genomic analyses. Plant Cell Physiol. 2013 Feb;54(2):e1.

[6] Mungall CJ, Torniai C, Gkoutos GV, Lewis SE, Haendel MA. Uberon, an integrative multi-species anatomy ontology. Genome Biol. 2012 Jan;13(1):R5.

[7] Buttigieg PL, Pafilis E, Lewis SE, Schildhauer MP, Walls RL, Mungall CJ. The environment ontology in 2016: bridging domains with increased scope, semantic density, and interoperation. J Biomed Semantics. 2016 Sep;7(1):57.

[8] Devare M, Aubert C, Laporte MA, Valette L, Arnaud E, Buttigieg PL. Data-driven Agricultural Research for Development: A Need for Data Harmonization Via Semantics. In: ICBO/BioCreative; 2016. Available from: https://api.semanticscholar.org/CorpusID:41553233.

[9] Dooley DM, Griffiths EJ, Gosal GS, Buttigieg PL, Hoehndorf R, Lange MC, et al. FoodOn: a harmonized food ontology to increase global food traceability, quality control and data integration. npj Sci Food. 2018 Dec;2(1):23. Available from: https://www.nature.com/articles/s41538-018-0032-6.

[10] Schaab GL. Relational Ontology. In: Runehov ALC, Oviedo L, editors. Encyclopedia of Sciences and Religions. Dordrecht: Springer Netherlands; 2013. p. 1974-5. Available from: http://link.springer.com/10.1007/978-1-4020-8265-8_847.

[11] National research institute for agriculture food and environment, National research institute for agriculture food and environment. Thésaurus INRAE. Recherche Data Gouv; 2021. Available from: https://entrepot.recherche.data.gouv.fr/citation?persistentId=doi:10.15454/J8GANU.

[12] Bernard M, Lecoeur A, Coville JL, Bruneau N, Jardet D, Lagarrigue S, et al. Relationship between feed efficiency and gut microbiota in laying chickens under contrasting feeding conditions. Scientific Reports. 2024 Apr;14(1):8210. Available from: https://doi.org/10.1038/s41598-024-58374-3.

ShareFAIR-KG, a centralised knowledge base of scientific workflows

Marie SCHMIT[1], Melvin Selim ATAY[2], Khalid BELHAJJAME[3], Ulysse LE CLANCHE[4], Emmanuel COQUERY[5], Olivier DAMERON[4], Fabien DUCHATEAU[5], Alban GAIGNARD[6,7], Mouna EL GARB[5], Jaffar GURA[3], Nicolas LUMINEAU[5], George MARCHMENT[8], Camille MAUMET[2], Clémence SEBE[8], Frédéric LEMOINE[1], Sarah COHEN-BOULAKIA[8] and Hervé MÉNAGER[1,7]

[1] Institut Pasteur, Université Paris Cité, Bioinformatics of Biostatistics Hub, F-75015 Paris, France
[2] Univ Rennes, CNRS, Inria, Inserm, IRISA UMR 6074, EMPENN — ERL U 1228, F-35000 Rennes, France
[3] LAMSADE, PSL, Paris Dauphine University, 75016 Paris, France
[4] Université Rennes, Inria, CNRS, IRISA—UMR 6074, Rennes 35000, France
[5] Université Lyon 1, INSA Lyon, CNRS, LIRIS, UMR 5205, Lyon, France
[6] Nantes Université, CNRS, INSERM, l'Institut du Thorax, F-44000 Nantes, France
[7] IFB-core, Institut Français de Bioinformatique (IFB), CNRS, INSERM, INRAE, CEA, 94800 Villejuif, France
[8] Université Paris-Saclay, CNRS, Laboratoire Interdisciplinaire des Sciences du Numérique, 91405, Orsay, France

Corresponding author: herve.menager@pasteur.fr

Keywords Knowledge graphs, workflows, FAIR, SPARQL, metadata

Abstract *The emergence of high-throughput technologies in the Life Sciences has led to increasingly complex data analysis pipelines. Ensuring their transparency, reusability, and reproducibility requires adherence to the FAIR principles, which motivated the development of the ShareFAIR project. ShareFAIR aims to build a framework to make workflows more shareable, understandable, and reusable by improving their description through structure, provenance, and annotations.*

The ShareFAIR architecture consists of several modules addressing workflow structure extraction, provenance, annotation, and querying, with applications ranging from genomics to neuroimaging. At its core, this system depends on ShareFAIR-KG, a centralized and queryable knowledge graph (KG) that integrates all workflow components and tool metadata at various levels of granularity. Populated through automated harvesting and normalized into a graph-based representation, it consolidates the results and software produced in the project. With the help of ShareFAIR-KG, users can for instance research workflows implementing specific analytical steps, obtain the software packages used or wrapped in workflows or query citation and licensing information.

Introduction

During the last decade, an unprecedented amount of multi-scale data has been generated in the field of Life Sciences. While it offers new opportunities for biological discovery, its volume and complexity have introduced significant computational challenges. Notably, the difficulty in reproducing data analysis pipelines has become prevalent, eroding trust in scientific results and limiting the reuse of established analysis pipelines [1]. This issue is particularly relevant in bioinformatics and extends to the neuroimaging field [2].

Scientific workflows were developed to address these limitations, offering scalability, automation and reproducibility [3]. However, workflows alone are insufficient: the application of FAIR principles (Findability, Accessibility, Interoperability and Reuse) [4,5,6,7] to these pipelines is increasingly recognised

as essential for ensuring transparency, quality and reliability of computational results [4]. Other authors have highlighted the importance to develop dedicated best practices for research software that also apply to workflows [8].

The ShareFAIR project arose from this need to make scientific workflows FAIR, easier to share and to reproduce. It seeks to enhance the reliability of both datasets (e.g. finely annotated raw data and analysis results) and analysis protocols by explicitly tracking provenance relationship between data results and analytical methods. To achieve this objective, the ShareFAIR architecture includes multiple modules for the annotation, query, extraction and representation of workflow structure and execution traces. The basis of this architecture is a knowledge graph of annotated workflows from different languages and sources, gathering and allowing to query data generated in those modules. This knowledge base facilitates the discovery of workflows with their scientific context, structure and execution traces.

In this work, we present ShareFAIR-KG, a knowledge-graph of normalised FAIR workflow metadata. It leverages existing state-of-the-art and community adopted standards to link the structural representation of scientific pipelines with their execution traces, software components and their parameters. Although there are several initiatives for sharing bioinformatics workflows and software such as the WorkflowHub directory [9] or the bio.tools registry [10,11], the aim of ShareFAIR-KG is to bring them together by adding a semantic layer on top of them. Thus, ShareFAIR-KG automatically harvests metadata from multiple existing repositories and enables their discovery through complex querying. The knowlegde base supports several use cases. Researchers can explore workflows implementing specific analytical steps, facilitating the location and comparison of workflows across all supported languages. Tool developers can query software packages used or wrapped in workflows, for instance to guide the creation of new wrappers. Finally, data managers and curators can evaluate workflow FAIR compliance by querying licensing or citation information. As the groundwork of the ShareFAIR architecture, ShareFAIR-KG integrates functionalities and results from ongoing collaborative development across multiple modules of the ShareFAIR project.

This paper is organised as follows. First, we present the standards selected for workflow representation, that build the semantics of ShareFAIR-KG. Next, we explain how the knowledge graph is deployed and populated, showing examples of use cases to demonstrate its query-ability and how it explains workflows at every granularity level. Finally, we detail how the knowledge base serves as the integration layer for other modules of the ShareFAIR project, and how we plan to improve this joint work to expand its content and capabilities.

Scientific workflow and tool metadata standards

Comprehensive workflow description requires a representation model with sufficient granularity, as workflows can be studied at different scales: (i) the workflows and their scientific domains at the highest level, (ii) then their processes, either nested subworkflows or computational units calling a (iv) software with its parameters. Those components and the data-flow between them must also be considered from both the prospective (specification) and retrospective (execution traces) provenance perspective.

This representation is further subject to additional critical requirements. First, it must integrate existing ontologies, schemas or vocabularies while avoiding to redefine yet another standard. Then, it must facilitate metadata generation and exchange by adopting interoperable and reusable standards that are widely adopted by the community and already integrated by existing frameworks.

The ShareFAIR-KG semantic model [12] complies with those constraints and provides (i) scientific context and metadata for workflows and their constituent software and steps, using domain ontologies such as EDAM [13]; (ii) unique software referencing via bio.tools [10,11] identifiers; and (iii) prospective and retrospective provenance representation throught Provenance Run Crate [14], an RO-Crate profile extending WorkflowHub's workflow profile (describing workflow files and metadata), Process Run Crate (software used for file creation) and Workflow Run Crate (workflow execution by a Workflow Management System or WMS).

Populating and querying ShareFAIR-KG

ShareFAIR-KG is a knowledge base of workflow metadata following this representation model. Queryable with SPARQL, it is deployed in a Jena Fuseki server, inside a Docker container providing isolation and reproducibility.

The knowledge base integrates (i) EDAM domain-specific ontology for annotation, (ii) bio.tools registry for software identification and as a source of pre-existing EDAM annotations, (iii) WorkflowHub general metadata on workflows, (iv) metadata on the prospective and retrospective provenance of workflows from various WMS and sources, including: WorkflowHub, nf-core, Common Workflow Language (CWL) [15,16] and the *Intergalactic Workflow Commission* (IWC) library. We chose those registries and libraries for their wide-spread adoption by the community and their alignment with FAIR principles, leveraging open registries (e.g. bio.tools) and standards (e.g. RO-Crate).

ShareFAIR-KG currently contains the prospective provenance of 49 Nextflow [17] and 73 Common Workflow Language (CWL) workflows; and both types of provenance for 102 Galaxy[18] workflows[18] executed on Galaxy server. This dataset spans workflows covering very diverse domains of bioinformatics, from genome assembly to metabolite profiling. Querying this knowledge graph with SPARQL yields deeper insights into these workflows, as shown in the following use cases (further detailed in [12]):

- **Tool adoption** An example query is the search for the most frequently used tools[19] from bio.tools among workflows in the knowledge base, with their EDAM operations. Querying ShareFAIR-KG reveals that the top two by frequency are MultiQC [19] and dada2 [20], performing the operations sequencing quality control and *DNA barcoding*, *variant calling* (Table 1).
- **Workflow using specific tools** Knowing that MultiQC is the most used tool amongst workflows in ShareFAIR-KG, the workflows using it can be queried. For example, the Galaxy workflows *Genome Assembly from Hifi reads - VGP3* and *Scaffolding with Hi-C data VGP8* both use MultiQC.
- **Workflow input format** Another example of query involves retrieving workflow input and output data. For instance, the inputs of the CWL workflow *gatk4W* have the formats FASTA and FASTQ.

[18] Workflows were selected based on open-source licencing and technical compatibility with our extraction tools.
[19] Tools that are the most frequently used in workflows, not the ones which have been run the most overall through these workflows.

– **Workflow tools output formats** The last query example infers workflow outputs from bio.tools EDAM annotations. Annotations of the bio.tools software executed in the Galaxy workflow *ATAC-seq* [21] indicate that it generates BED, WIG, TSV, HTML, JSON, YAML and CSV data.

Tab. 1. Results of the SPARQL query retrieving most-used bio.tools tools, and their integration frequency in steps (# workflow steps). The names of the tools are from bio.tools. The namespace edam corresponds to http://edamontology.org/.

Tool name	bio.tools identifier	EDAM operation	EDAM operation label	# workflow steps
MultiQC	https://bio.tools/multiqc	edam:operation_3218, edam:operation_2428	Sequencing quality control, Validation	18
dada2	https://bio.tools/dada2	edam:operation_3200, edam:operation_3227	DNA barcoding, Variant calling	12
Bandage	https://bio.tools/bandage	edam:operation_3184	Sequence assembly visualisation	5

Current and Future Integration within the ShareFAIR Project

ShareFAIR-KG constitutes the foundational infrastructure of the ShareFAIR architecture, centralising output from other modules and providing data for query and representation purposes to others (Figure 1). While it already integrates several of their functionalities, ongoing efforts aim to strengthen this integration and consolidate the ShareFAIR framework. This collaboration will address key limitations by enhancing annotation depth and quality as well as expanding workflow coverage.

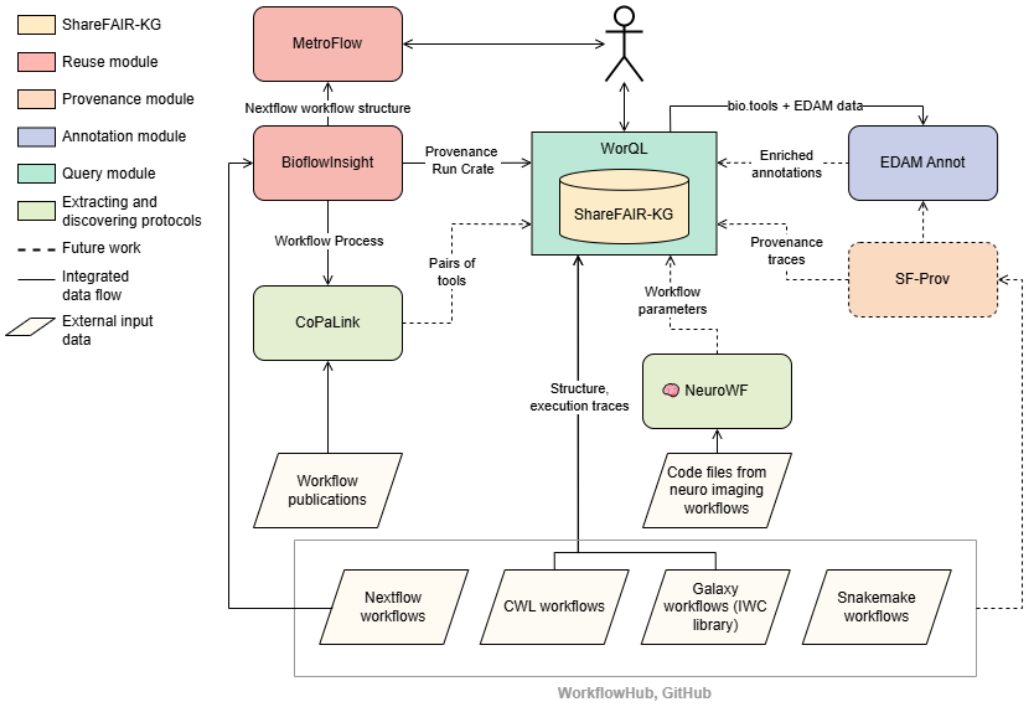


Fig. 1. ShareFAIR architecture with the software developed in different modules and the data flowing between them (type of data is specific on the arrows). Solid arrows represent current integration, dotted arrows show work in progress.

The Reuse Module aims to extract and visualise workflow structure. As part of this module, BioflowInsight [22] was developed to capture Nextflow workflow structures and generate their Provenance Run Crate metadata within ShareFAIR-KG. Additionally, MetroFlow [23] provides a unique visualization of Nextflow workflows, representing them as intuitive metro maps.

The Provenance Module ingests and exploits provenance traces captured from workflow executions. Only a limited number of WMS supports provenance capture, such as Galaxy and Nextflow with the plugin nf-prov. This module currently aligns nf-prov provenance with prospective specifications. In the short-term, our road map includes the verification and repair of the completeness of nf-prov outputs. In the longer term, we investigate multi-granular representations of provenance, including fine-grained execution traces. Both the repair and the refined provenance derivations (at different levels of granularity) will be reflected in ShareFAIR-KG.

The Annotation Module aims at inferring missing metadata to better annotate and explain scientific pipelines, as well as analyzed data. Bio.tools metadata and the EDAM ontology from ShareFAIR-KG are at the core of the EDAMAnnot [24] toolbox and were used to investigate metadata quality in terms of information content. Neuro-symbolic approaches leveraging both retrieval augmented generation and knowledge graphs are under investigation to i) increase the quality of bioinformatics tools and workflow metadata and ii) annotate analysed data for increase potential reuse.

The Query Module leverages BioFlow-Ontology [25], a comprehensive workflow representation model, to develop WorQL, a dedicated query language for workflow interrogation. WorQL employs *query relaxation mechanisms*, such that if a query yields no results, it is automatically reformulated to return meaningful matches [26]. This enhances ShareFAIR-KG accessibility, enabling users from diverse backgrounds to efficiently query workflows and fostering broader adoption. Bioflow-Ontology and the ShareFAIR-KG representation model serve complementary purposes: one extends existing standards for seamless metadata generation and interoperability, the other optimises workflow querying. The integration of this module with ShareFAIR-KG will leverage inference rules and reasoning to align both data representations.

Linking bioinformatics tools across publication and workflow code Workflow descriptions in publications and actual code often diverge. For example, the same tool may appear under different names, or can be omitted from the publication. To address this, CoPaLink [27] has been developed to extract all bioinformatics tool names from both sources (publications and code). This allows users to compare tools detected in each source and identify tools that appear in only one source. CoPaLink will be integrated in ShareFAIR-KG.

Neuroimaging Workflow Integration (NeuroWF), currently in development Besides bioinformatics workflows and provenance tools, ShareFAIR-KG will enable querying and exploring the collection of open source neuroimaging workflows. In particular, we will generate Provenance Run Crate metadata to describe functional magnetic resonance imaging (task-fMRI) pipelines of the most commonly used software tools. By using the previously described modules, NeuroWF aims to enable parameter metadata [28] querying for further analysis of neuroimaging workflows in a reproducible manner.

Conclusion

This work paves the way for FAIR sharing of scientific workflows. It leverages the results and functionalities developed in the ShareFAIR project modules to gather and normalise workflow annotations and metadata from multiple sources and languages, providing details on their scientific context, structure and execution traces to facilitate workflow discovery, understanding, sharing and reuse.

ShareFAIR-KG still faces specific constraints, mainly related to the quality and completeness of annotations and workflow descriptions. Query capabilities are currently limited by the variable depth and coverage of workflow description. Additionally, the knowledge graph is rebuilt from scratch during each update rather than maintained incrementally with workflow versioning support, limiting its scalability for evolving pipelines.

Future work will first address the lack of complete workflow annotations by finalising the integration with the other ShareFAIR modules, to integrate a larger diversity of standards and increase annotation quality, for instance by enriching bio.tools annotations. Furthermore, we plan to increase the scale of workflow integration to encompass neuroimaging workflows, while expanding the coverage of execution traces beyond Galaxy workflows. Joint efforts will also focus on improving query accessibility and user-friendliness. These advances will establish the technical foundation for the ShareFAIR framework.

Availability and implementation

All resources are openly available: **ShareFAIR-KG** with examples of queries and competency questions (GitLab, v1.0.0, GNUGPLv3.0, `swh:1:dir:4e761e3`), **EDAMAnnot** (GitHub, GNUGPLv3.0, `swh:1:dir:bf911b6`), **BioFlowInsight** (GitLab, GNUAGPLv3.0), **NeuroWF** (GitLab, MIT License, `swh:1:dir:467be46`), and **CoPaLink** (tool and experiment repositories, GNUAGPLv3.0) are hosted on GitLab and GitHub with persistent Software Heritage archives; the **knowledge graph content** (Zenodo, DOI:10.5281/zenodo.17737888) and **BioFlow-Ontology** (Zenodo, v7, CC-BY4.0) are available on Zenodo.

Acknowledgments

This work was supported in part by the National Research Agency under the France 2030 program, with reference to ANR-22-PESN-0007 (ShareFAIR project).

Generative AI Statement

During the preparation of this work, the authors used Kimi K2 (Moonshot AI) and Claude Sonnet 4.5 to improve textual clarity through reformulation, as well as to verify grammar and spelling. After using this tool/service, the authors reviewed and edited the content as needed and take full responsibility for the content of the publication.

Citations

References

[1] Cohen-Boulakia S, Belhajjame K, Collin O, Chopard J, Froidevaux C, Gaignard A, et al. Scientific Workflows for Computational Reproducibility in the Life Sciences: Status, Challenges and Opportunities. Future Generation Computer Systems. 2017 Oct;75:284-98.

[2] Botvinik-Nezer R, Holzmeister F, Camerer CF, Dreber A, Huber J, Johannesson M, et al. Variability in the analysis of a single neuroimaging dataset by many teams. Nature. 2020 Jun;582(7810):84-8. Available from: https://doi.org/10.1038/s41586-020-2314-9.

[3] Djaffardjy M, Marchment G, Sebe C, Blanchet R, Belhajjame K, Gaignard A, et al. Developing and reusing bioinformatics data analysis pipelines using scientific workflow systems. Computational and Structural Biotechnology Journal. 2023 Jan;21:2075-85. Available from: https://www.sciencedirect.com/science/article/pii/S2001037023001010.

[4] Goble C, Cohen-Boulakia S, Soiland-Reyes S, Garijo D, Gil Y, Crusoe MR, et al. FAIR Computational Workflows. Data Intelligence. 2020;2(1-2):108-21. Available from: https://doi.org/10.1162/dint_a_00033.
[5] Wilkinson SR, Aloqalaa M, Belhajjame K, Crusoe MR, de Paula Kinoshita B, Gadelha L, et al. Applying the FAIR principles to computational workflows. Scientific Data. 2025;12(1):328.
[6] Wilkinson MD, Dumontier M, Aalbersberg IJ, Appleton G, Axton M, Baak A, et al. The FAIR Guiding Principles for Scientific Data Management and Stewardship. Scientific Data. 2016 Mar;3(1):160018.
[7] Barker M, Chue Hong NP, Katz DS, Lamprecht AL, Martinez-Ortiz C, Psomopoulos F, et al. Introducing the FAIR Principles for Research Software. Scientific Data. 2022 Oct;9(1):622.
[8] Di Cosmo R, Granger S, Hinsen K, Jullien N, Le Berre D, Louvet V, et al. CODE beyond FAIR: a roadmap for reusable research software. Scientific Data. 2025;In Press.
[9] Gustafsson OJR, Wilkinson SR, Bacall F, Pireddu L, Soiland-Reyes S, Leo S, et al.. WorkflowHub: a registry for computational workflows. arXiv; 2024. ArXiv:2410.06941 [cs]. Available from: http://arxiv.org/abs/2410.06941.
[10] 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 Research. 2016 Jan;44(D1):D38-47. Available from: https://doi.org/10.1093/nar/gkv1116.
[11] Ison J, Ienasescu H, Chmura P, Rydza E, Ménager H, Kalaš M, et al. The bio.tools registry of software tools and data resources for the life sciences. Genome Biology. 2019 Aug;20(1):164. Available from: https://doi.org/10.1186/s13059-019-1772-6.
[12] Schmit M, Le Clanche U, Marchment G, Cohen-Boulakia S, Dameron O, Gaignard A, et al. A Standards-Based Knowledge Graph that Bridges Scientific Workflows, Run-Time Provenance, and Tool Registries. In: SWAT4HCLS 2026. Amsterdam, Netherlands; 2026. Available from: https://hal.science/hal-05517640.
[13] Gaignard A, Skaf-Molli H, Belhajjame K. Findable and reusable workflow data products: A genomic workflow case study. Semantic Web. 2020;11(5):751-63. Available from: https://journals.sagepub.com/doi/abs/10.3233/SW-200374.
[14] Leo S, Crusoe MR, Rodríguez-Navas L, Sirvent R, Kanitz A, De Geest P, et al.. Recording Provenance of Workflow Runs with RO-Crate. arXiv; 2024.
[15] Amstutz P, Crusoe MR, Tijanić N, Chapman B, Chilton J, Heuer M, et al.. Common Workflow Language, v1.0; 2016. Publisher: figshare. Available from: https://research.manchester.ac.uk/en/publications/common-workflow-language-v10/.
[16] Crusoe MR, Abeln S, Iosup A, Amstutz P, Chilton J, Tijanić N, et al. Methods included: standardizing computational reuse and portability with the common workflow language. Communications of the ACM. 2022;65(6):54-63.
[17] 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.
[18] The Galaxy Community. The Galaxy Platform for Accessible, Reproducible and Collaborative Biomedical Analyses: 2022 Update. Nucleic Acids Research. 2022 Jul;50(W1):W345-51.
[19] Ewels P, Magnusson M, Lundin S, Käller M. MultiQC: summarize analysis results for multiple tools and samples in a single report. Bioinformatics. 2016 06;32(19):3047-8. Available from: https://doi.org/10.1093/bioinformatics/btw354.
[20] Callahan BJ, McMurdie PJ, Rosen MJ, Han AW, Johnson AJA, Holmes SP. DADA2: High-resolution sample inference from Illumina amplicon data. Nature methods. 2016;13(7):581-3.
[21] Lopez-Delisle L. github.com/iwc-workflows/atacseq/main. Zenodo; 2025. Available from: https://doi.org/10.5281/zenodo.15078780.
[22] Marchment G, Brancotte B, Schmit M, Lemoine F, Cohen-Boulakia S. BioFlow-Insight: facilitating reuse of Nextflow workflows with structure reconstruction and visualization. NAR Genomics and Bioinformatics. 2024 08;6(3):lqae092. Available from: https://doi.org/10.1093/nargab/lqae092.

[23] Marchment G, Brancotte B, Cohen-Boulakia S, Gura J, Lemoine F. MetroFlow : Generating Interactive Metro-Maps from Workflow Code. Nextflow Summit 2025. 2025. Available from: https://www.youtube.com/watch?v=mgR7wQIGUDo.

[24] Le Clanche U, Cohen-Boulakia S, Cunff YL, Dameron O, Gaignard A. Assessing bioinformatics software annotations: bio.tools case-study. In: Proceedings JOBIM. Bordeaux & Online, France; 2025. p. 1-7. Available from: https://inria.hal.science/hal-05096849.

[25] El Garb M, Coquery E, Duchateau F, Lumineau N. Improving reproducibility in bioinformatics workflows with BioFlow-Model. In: ACM Conference on Reproducibility and Replicability (ACM REP '25),. Vancouver, Canada: ACM; 2025. Available from: https://hal.science/hal-05240352.

[26] Fakih G, Serrano-Alvarado P. A survey on SPARQL query relaxation under the lens of RDF reification. Semantic Web. 2024;15(6):2507-54.

[27] Sebe C, Ferret O, Névéol A, Esmailoghli M, Leser U, Cohen-Boulakia S. Supporting Workflow Reproducibility by Linking Bioinformatics Tools across Papers and Executable Code. arXiv; 2026. ArXiv:2603.08195 [cs]. Available from: http://arxiv.org/abs/2603.08195.

[28] Atay MS, Clenet B, Bannier E, Maumet C. Real-world parameter preferences in task-fMRI preprocessing pipelines. In: OHBM 2026 - Annual Meeting of the Organization for Human Brain Mapping. Bordeaux, France: OHBM; 2026. Available from: https://inria.hal.science/hal-05545909.

Demos

Depictio: an open-source platform for building interactive dashboards from bioinformatics workflow outputs

Demo

Thomas Weber [1], ***Jan Korbel*** [1]

1. *European Molecular Biology Laboratory*

Abstract

Bioinformatics relies on standardized workflows executed through engines such as Nextflow, Snakemake, or Galaxy, producing large, heterogeneous outputs across multiple runs. Yet researchers still lack a unified way to aggregate, explore, and share these results interactively, without writing custom visualization code for each project.

We present Depictio, an open-source web platform that transforms bioinformatics workflow outputs into interactive, shareable dashboards. Depictio supports three main user profiles. First, researchers can build dashboards from scratch by defining a project, ingesting data through a command-line interface, and composing visualizations via drag-and-drop UI or a YAML configuration. Second, users with completed analyses can turn their results into publication-ready companion apps, shareable with collaborators through simplified deployment and permalinks. Third, community-driven dashboard templates (nf-core for instance) allow users to instantly populate dashboards from pipeline outputs. A single CLI command automatically ingests data and generates a ready-to-explore dashboard with no manual setup required.

A key strength of Depictio is its rich interactivity. Multi-tab dashboards organize complex analyses into logical views, each composed from a library of components including figures, tables, metric cards, embedded MultiQC reports, image galleries, and geospatial maps. Cross-component filtering operates through multiple interaction modes: lasso and box-select on scatter plots, click-to-select on individual data points, row selection in tables, and marker selection on maps. These interactions propagate in real time across all linked components, enabling fluid exploratory analysis. Users can further refine their views with dropdown filters, sliders, and date pickers.

Depictio is cloud-ready, deployable via Docker Compose or Kubernetes, and is available as a hosted application on SciLifeLab Serve (https://serve.scilifelab.se) through a collaboration with SciLifeLab Data Centre. A live demo is accessible at https://demo.depictio.embl.org. During the demonstration, we will showcase dashboard creation, interactive data exploration, and template-based instant setup from nf-core pipeline results.

URL

https://depictio.github.io/depictio-docs/

Gaston 2, a C++ library and an R package for large-scale genotype data

Demo

Hervé Perdry [1], ***Juliette Meyniel*** [1]

1. *Université Paris-Saclay, UVSQ, Inserm, CESP*

Abstract

In 2015, the first version of the gaston R package was released on CRAN (authors Claire Dandine-Roulland, Hervé Perdry). It was designed to handle genetic datasets up to several thousand individuals, and to a million variants. This package became quite successful, with circa 50,000 downloads per year on the Rstudio mirror, ranking in the top 10 percent of CRAN packages (figures obtained with packageRank).

However, in the last ten years, the size of genetic datasets increased dramatically. In-memory computation became unfeasible, not only for storing genetic data, but for computations deriving from those, for example Genetic Relationship Matrix (GRM).

Gaston 2 (or more precisely, gaston2) is a completely new library, with two components:

- A header-only C++ library, allowing for easy interfacing with other R packages or even with other languages
- An R package, providing a user-friendly interface to the C++ objects in the versatile R ecosystem.

Features

- Data can be loaded in memory or manipulated on disk (memory-mapped objects). In parallel of the development of gaston2, we developed houba (released on CRAN) for manipulating memory-mapped numeric matrices. In addition, gaston2 has classes for manipulating memory-mapped data in the plink format.
- While gaston focused on genotype data, gaston2 will allow the manipulation of dosage data.
- The C++ functions are templated allowing for single (float) or double precision calculations, to optimize memory usage and (marginally) speed. Many functions are parallelized using openMP.

Demonstration at JOBIM 2026

We propose to demonstrate the R package, showing how to perform a quality control procedure and a genome-wide association study. A quick glance to the C++ library can be proposed if the public shows interest.

URL

https://github.com/genostats/gaston2

IFB-Biosphère: Open access to adaptable computing resources within reproducible environments

Demo

Matis Zouari [1], ***Mateo Boudet*** [2], ***Guillaume Brysbaert*** [3], ***Micael Calvas*** [4], ***Stephane Delmotte*** [5], ***Hervé Gilquin*** [4], ***Nadia Goué*** [6], ***Jean François Guillaume*** [7], ***Antoine Mahul*** [8], ***Jérôme Pansanel*** [9], ***Bruno Spataro*** [5], ***Cyrille Toulet*** [10], ***Christophe Blanchet*** [11]

1. *IFB-core CNRS UAR3601,* ***2.*** *GenOuest, Univ Rennes, Inria, CNRS, IRISA, 35000, Rennes, France,* ***3.*** *Univ. Lille, CNRS UMR 8576-UGSF-Unité de Glycobiologie Structurale et Fonctionnelle, 59000 Lille, France,* ***4.*** *CBPsmn, Centre Blaise Pascal de simulation et modélisation numérique, Ecole Normale Supérieure de Lyon, France,* ***5.*** *Universite Claude Bernard Lyon 1, LBBE, UMR 5558, CNRS, VetAgro Sup, Villeurbanne 69622, France,* ***6.*** *Plateforme AuBi, Mésocentre Clermont-Auvergne, Université de Clermont-Ferrand, Aubière, France,* ***7.*** *GLiCID, BiRD, UMS BioCore (Inserm US16 et UAR CNRS 3556), UFR Médecine et Techniques Médicales, Nantes Université, CHU Nantes, France,* ***8.*** *Mésocentre Clermont-Auvergne, Université de Clermont-Ferrand, Aubière, France,* ***9.*** *Université de Strasbourg, CNRS, IPHC UMR 7178, F-67000 Strasbourg, France,* ***10.*** *Mésocentre régional, Université de Lille, France,* ***11.*** *IFB-core, Institut Français de Bioinformatique (IFB), CNRS, INSERM, INRAE, CEA, 94800 Villejuif, France.*

Abstract

Today, bioinformatics has to deal with an ever growing set of tools for computation, visualisation and analysis. Each tool's use depends on the field of study and the object of the research (genomics, transcriptomics, metagenomics), and bioinformaticians have their own preferences as to which environment they choose to perform analysis with. Therefore, facilitating **access to computing resources** compatible with a **wide variety of tools** to choose from can greatly accelerate biological research as a whole.

In this context, the Biosphere web portal, hosted by the IFB (**French Bioinformatics Institute**), gives open access to a wide catalog of **high-availability** environments with pre-installed tools. These environments are enclosed and allow users to run **reproducible** analysis, with mainstream technologies such as RStudio, Jupyter, Shell scripts, Ansible, workflows and packages. They are **easily adaptable**, enabling the installation of additional tools with minimal technical knowledge by any kind of user.

These environments can be used by **scientists** for their own research, and by teams in the **context of a project** involving a pre-definite set of tools. They can also be used within the framework of **training courses**, giving access to **students** to their own sandbox environment so that they can learn how to use specific tools.

The users get access to these computing resources via a **Cloud-based architecture**. The Biosphere web platform acts as the center-point between the users and the 8 partner centers (bioinformatics platforms and regional computing centers) which dedicate hardware to these environments. In this demo, we would like to show how potent **enclosed, available and adaptable** environments are regarding **reproducible** science, and to help potential users to benefit from the cloud bioinformatics services of Biosphere.

Biosphere's architecture is also undergoing evolutions and will soon integrate Kubernetes as a means to deploy containers as environments, giving even more flexibility for deploying environments.

URL

https://biosphere.france-bioinformatique.fr/catalogue

madbot, a metadata and data brokering online tool to ensure the adoption of standards and FAIR principles in an open science context

Demo

Imane MESSAK [1], ***Baptiste Rousseau*** [1], ***Elora Vigo*** [1], ***madbot working group*** [1], ***Hélène CHIAPELLO*** [1], ***Nadia Goué*** [2], ***Julien Seiler*** [1], ***Thomas Denecker*** [1]

1. *IFB-core, Institut Français de Bioinformatique (IFB), CNRS, INSERM, INRAE, CEA,* ***2.*** *Plateforme AuBi, Mésocentre Clermont-Auvergne, UCA*

Abstract

madbot (metadata and data brokering online tool) is a tool that helps researchers manage and share scientific data more easily. Developed by the Institut Français de Bioinformatique (IFB), madbot addresses the growing challenges of ensuring that research data are accessible, reusable, and well described as data volumes continue to increase. Existing tools often address only parts of this process and lack automation, standardization, or flexibility. madbot provides a comprehensive solution that follows international data standards and automates data organization and metadata curation, saving time and effort for researchers. The tool directly connects to major repositories such as European Nucleotide Archive (EBI-ENA) or Zenodo, enabling seamless data submission through an intuitive interface that does not require technical expertise. Behind the scenes, madbot ensures data consistency during the transfer to repositories, validates metadata with metadata referential and ontologies, and supports high-quality descriptions. Its extensible architecture allows integration with various data storage systems, repositories, and metadata standards. By simplifying data publication, madbot lowers barriers to data sharing and promotes open science within the global research community. madbot is available here : https://madbot.france-bioinformatique.fr/. During the demo, we will present the key concepts of madbot and illustrate them through a use case: (1) linking data present on the IFB core cluster with madbot, (2) describing this data using controlled and standardized metadata, and (3) submit data and metadata to the ENA reference database.

URL

https://madbot.france-bioinformatique.fr/

MetroFlow: automatic, interactive metro-map visualisation for enhancingtransparency and comprehensibility of Nextflow workflows

Demo

George Marchment [1], ***Bryan Brancotte*** [2], ***Jaffar Gura*** [3], ***Frédéric Lemoine*** [4], ***Sarah Cohen-Boulakia*** [1]

1. *Université Paris-Saclay, CNRS, Laboratoire Interdisciplinaire des Sciences du Numérique,* ***2.*** *Institut Pasteur, Université Paris Cité, Bioinformatics and Biostatistics Hub, Paris, France,* ***3.*** *LAMSADE, Université Paris Dauphine,* ***4.*** *Institut Pasteur, Univ. Paris Cité, National Reference Center for Respiratory Viruses, Bioinformatics and Biostatistics Hub*

Abstract

Research based on computational methods continues to grow rapidly, reflecting the increasing dependency on high-throughput computational pipelines for data-intensive scientific analysis. Scientific workflows systems such as Nextflow [1] have emerged as tools to improve computational reproducibility and scalability, and are now widely shared on public platforms, creating a large collection of reusable resources [2]. Cumulative science is more important than ever, enabling researchers to compare, refine, and build upon prior methods is important. The ability to understand and trust existing workflows is essential. Yet workflows describing complex data analyses are often very large and highly modular, making them difficult to interpret and reuse without clear documentation and accessible representations.

The nf-core community shows how high-quality visual documentation (based on metro-map representations) facilitates reuse [3]. However, such visual documentation is produced manually by developers. Maintaining nf-core workflows requires substantial and coordinated effort. Keeping documentation aligned with such a complex, evolving code thus remains a significant challenge.

To bridge this gap, we introduce MetroFlow, a tool designed to make workflows more transparent and easier to understand. MetroFlow automatically condenses thousands of lines of Nextflow code into an interactive, multi-level structural map. It enables users to explore the workflow logic by visualising (i) workflow steps and their associated code, (ii) the conditions that determine whether a path is executed, and (iii) subworkflows in isolation. In addition, users can modify the workflow layout by (iv) rearranging its components and (v) collapsing subworkflows into a single node, thereby reducing the complexity of the overall representation. These functionalities allow users to navigate the main analysis steps at various levels of granularity, anticipate the workflow's execution, and progressively explore its components. This approach reduces the documentation burden while improving workflow code interpretability for developers, users, and reviewers, making workflow logic clearer, more accessible, and easier to reuse.

URL

https://metroflow.pasteur.cloud/

MOAL - MULTI-OMIC ANALYSIS AT LAB A R PACKAGE TO IMPROVE THE ACCESSIBILITY AND REPRODUCIBILITY OF OMICS BIOANALYSIS

Demo

Florent Dumont [1]

1. *UMS IPSIT, Paris-Saclay University*

Abstract

Exploiting omic data has become usual in many medical and biological research laboratories. The increasing number of technological platforms and their attached services, the lessening of costs, and the publication of raw and processed data impulse the creation of new standards in analysis methodology. In this context, numerus bioinformatics tools are available to make preprocessing and generate raw data matrix. The following step is normalization which aims to increase the signal-to-noise ratio and harmonize sample distributions to make differential analysis between groups depending on the experimental design.

Typically, **bioanalysis** defines downstream analysis steps including supervised and unsupervised discriminant analysis which generate some variable clusters from each data sets which subsequently need to be integrated to scientific knowledge. Such process currently named, functional enrichment analysis, is used to facilitate biological interpretation, curing statistical false positives, and also to create infographics to better understand and valorise the biological complexity of multi-omic data sets.

In order to better control analysis reproducibility, facilitate data exploration and open bioanalysis, we developed MOAL (Multi Omic Analysis at Lab) [1] a R package including an easy-to-use unique function starting from tabular files, the normalized data matrix and the sample metadata. The workflow automates most classical tasks and generates in one go results tables and interpretable graphics for both biostatics and functional analysis.

Install and test demo scripts from github and r-universe at https://github.com/fdumbioinfo/moal

[1] Dumont F, Ponsardin E, Bernadat G, Cohen-Kaminsky S. MOAL: Multi-Omic Analysis at Lab. A simplified methodology workflow to make reproducible omic bioanalysis. bioRxiv oct.2023. 10.1101/2023.10.17.562686, 10.5281/zenodo.10551966.

URL

github: https://github.com/fdumbioinfo/moal

r-universe: https://fdumbioinfo.r-universe.dev/moal

Pixitainer: frictionless apptainer image generation from a pixi workspace

Demo

Raphaël Ribes [1]

1. *ISDM, Univ Montpellier, CIRAD, INSERM, Montpellier, France*

Abstract

Modern bioinformatics relies heavily on workflow managers like Nextflow and Snakemake to process massive multi-omics datasets across High-Performance Computing (HPC) environments. To ensure reproducibility and portability for those workflows, containerization has become a standard practice, with Apptainer/Singularity being the preferred choice for HPC deployments. However, developing container images is often a tiresome, iterative process involving manual definition file creation, system-level dependency resolution, and heavy build times.

Simultaneously, pixi has brought a breath of fresh air to package management solutions by providing extremely fast, lockfile-based environments. Backed by the Conda ecosystem (including bioconda) and supporting automation via TOML manifests, pixi allows developers to quickly iterate on software stacks. Despite its speed during local development, deploying a pixi environment directly into an HPC pipeline remains challenging for keeping reproducibility without a proper containerization strategy.

I present Pixitainer, a command-line extension that bridges this gap by offering a frictionless, automated from a local pixi development environment to a production-ready Apptainer and Docker image. Pixitainer natively translates a project's manifest into an Apptainer/Docker container. Designed specifically for scientific pipelines, it generates minimal images that freeze exact locked dependencies, ensuring high reproducibility when executed on computing nodes. Scientists can define container specifics directly in their manifest files, circumventing the need to write Apptainer definition code or Dockerfile. Additionally, a seamless execution mode configures the container to automatically proxy commands through the isolated Pixi environment, allowing the image to act as a drop-in executable replacement for the raw tools.

URL

https://github.com/RaphaelRibes/pixitainer

SaVanache : Interactive Visualization of Pangenomic Diversity

Demo

Mourdas Mohamed [1], ***François Sabot*** [1]

1. *DIADE, équipe PANEEC IRD Montpellier 911 avenue d'Agropolis BP 64501 34394 Montpellier Cedex 5*

Abstract

SaVanache is an interactive visualization tool dedicated to exploring pangenomic diversity, whether through comparisons between individuals across multiple pangenomes or through the analysis of structural variations (SVs) within a single pangenome. It enables intuitive navigation through genomic rearrangements insertions, deletions, and inversions while providing a continuous zoom from the chromosomal to the nucleotide scale. Originally developed as a prototype within the framework of a doctoral thesis (Durant, 2022), Savanache has since been fully implemented and extended. Combining a high-performance indexing engine with a responsive web interface, it dynamically highlights affected regions and their genomic context. Compatible with the GFA graph format, it now supports both population-level and individual-level analyses, bridging large-scale pangenome graph data with biological interpretation in an efficient and scalable way.

URL

https://forge.ird.fr/diade/savanache/savanache

VCFProcessor: a complete toolbox for improved VCF file analysis

Demo

Thomas LUDWIG [1], ***Gaëlle Marenne*** [1], ***Emmanuelle Génin*** [1]

1. *INSERM UMR1078, Univ Brest, EFS, CHRU Brest : Génétique, Génomique Fonctionnelle et Biotechnologies*

Abstract

VCF is the prevailing file format to store genetic variants. While it is produced by most variant callers and can be processed by a large range of tools, it is not always straightforward to parse “by hand” using standard Unix tools (cut,grep,head,...) or even awk/python in order to execute particular operations. Numerous tools have been developed over the years to perform different tasks on VCF file. However, these tools are often too generalist or too specific to suit the needs of users from a given field.

Here we propose VCFProcessor, that has been developed to answer precise biological/population genetics questions. It is a command-lines/GUI tool kit that allows users to carry out a large selection of functions on VCF files such as annotations, transformations, filtering or analyses. VCFProcessor is able to exploit sample properties (e.g., phenotype or population), allowing to perform useful comparisons or filtering. It is also possible to plot analyses’ results on relevant graphs. Finally, users can easily extend this tool kit, by developing their own functions, tailored to answer specific questions about the data at hand. As it is the case with native ones, these user defined functions will benefit from multi-threading and parallel processing.

URL

https://lysine.univ-brest.fr/vcfprocessor

Virome@tlas-explorer: Putting the virosphere on the map

Demo

Luca Nesterenko [1], ***Elea Pauliat*** [1], ***Paul Tissot*** [1], ***Mélodie Fleury*** [1], ***Maël Rimeur*** [1], ***Stephane Delmotte*** [2], ***Romain Delunel*** [3], ***Julien DELLINGER*** [1], ***Caroline Leroux*** [4], ***Jérome Lejot*** [3], ***Romuald Marin*** [5], ***Matis Zouari*** [6], ***Christophe Blanchet*** [6], ***Dominique Guyot*** [1], ***Christine Oger*** [1], ***François Mialhe*** [3], ***Hussein Anani*** [7], ***Julien Barnier*** [8], ***Damien de Vienne*** [9], ***Laurence Josset*** [7], ***Jocelyn Turpin*** [4], ***Oldrich Navratil*** [3], ***Vincent Navratil*** [1]

1. *PRABI, Rhône-Alpes Bioinformatics Center, Université Lyon 1, 43 bd du 11 novembre 1918, Villeurbanne CEDEX 69622, France,* ***2.*** *Universite Claude Bernard Lyon 1, LBBE, UMR 5558, CNRS, VetAgro Sup, Villeurbanne 69622, France,* ***3.*** *CNRS 5600 EVS, Université Lumière Lyon 2,* ***4.*** *IVPC UMR754, INRAE, Universite Claude Bernard Lyon 1, EPHE, PSL Research University, 69007, Lyon, France,* ***5.*** *CNRS, UAR 3601 ; Institut Français de Bioinformatique, IFB-core, 7 rue Guy-Môquet, F-94800 Villejuif, France,* ***6.*** *IFB-core, Institut Français de Bioinformatique (IFB), CNRS, INSERM, INRAE, CEA, 94800 Villejuif, France.,* ***7.*** *Hospices civils de Lyon,* ***8.*** *LBBE,* ***9.*** *Laboratoire de Biométrie et Biologie Évolutive, Lyon, France*

Abstract

Improved preparedness for future pandemics relies on monitoring the virosphere of wildlife and domestic animals along with the environment they share with humans. This surveillance can only be implemented through the development of interdisciplinary One-Health consortia, with digital solutions designed to integrate and visualise massive heterogeneous metagenomic datasets simultaneously, at a global scale and in quasi real-time. The Virome@tlas project, initiated with that major goal, is now powered by a data lake cyberinfrastructure which has facilitated the generation of a high-quality dataset made of 50 million georeferenced viral sequences/biosamples enriched with taxonomy, virus-host relationships, geoenvironmental and spatio-temporal metadata.

This motivates the development of Virome@tlas-explorer, an open source and web-based interactive platform for the visualization and analysis of viral samples. Its aim is two-fold: firstly, it gives users access to the comprehensive datasets curated within the project, allowing their full exploration and exploitation. Secondly, the modularity and flexibility of the tool enable its use with custom datasets uploaded by the users, making it invaluable for researchers. One can choose between a basic or expert mode, reflecting this duality. The first allows intuitive exploration of the curated datasets via simple queries to identify samples associated with a given host, environment, matrix or virus, visualised as a table or through a cartographic or taxonomic representation. Conversely, the advanced tool, with a customisable interface, enables a more in-depth exploration of these datasets or those that the users may upload. Besides allowing advanced queries, it provides supplementary visualisations which can interactively be used to refine the user's focus, making it straightforward to limit the data exploration to e.g., a time interval or geographical region of interest.

The capabilities of the tool and its value for large-scale virome research will be presented, showcasing its features and potential via several relevant use cases.

URL

https://viromeatlas.univ-lyon1.fr/ui (password: Trojan_horse37866)

ViromeChat-AI: a conversational interface to explore viral metagenomic data in the Virome@tlas project

Demo

Romuald Marin [1], ***Elea Pauliat*** [2], ***Paul Tissot*** [2], ***Mélodie Fleury*** [2], ***Luca Nesterenko*** [3], ***Oldrich Navratil*** [4], ***Vincent Navratil*** [2], ***Christophe Blanchet*** [5]

1. *IFB-core, Institut Français de Bioinformatique (IFB), CNRS, INSERM, INRAE, CEA,* ***2.*** *PRABI, Rhône-Alpes Bioinformatics Center, Université Lyon 1, 43 bd du 11 novembre 1918, Villeurbanne CEDEX 69622, France,* ***3.*** *Université Claude Bernard Lyon 1,* ***4.*** *CNRS 5600 EVS, Université Lumière Lyon 2,* ***5.*** *IFB-core, Institut Français de Bioinformatique (IFB), CNRS, INSERM, INRAE, CEA, 94800 Villejuif, France.*

Abstract

Recent pandemics have highlighted the connections between human, animal and environmental health. A comprehensive analysis of virus-host-ecosystem relationships at global scales is therefore essential for better characterizing the virosphere and the emergence of new viral pathogens. The Virome@tlas project is a cloud-based platform for global virus surveillance, integrating harmonized geographic, taxonomic (virus/host), and environmental metadata with viral sequences from public archives (NCBI, GenBank, SRA, GSA).

In a highly pluridisciplinary context and within a perspective of open science aligned with FAIR principles, one of the major goals is to make these data available to a wide range of scientists, not only virologists or data scientists. An AI-based conversation assistant could facilitate the exploration and analysis of virus-host-environment interactions across a wide range of disciplines and questions (e.g., ecology, health sciences, biogeography), and foster innovative ideas and new research hypotheses related to viral emergence and the global surveillance of socio-ecosystem.

We developed ViromeChat-AI, an LLM agent web interface designed to i) facilitate access to harmonized metadata related to viruses from the Virome@tlas datalake infrastructure and external information sources such as Wikipedia and PubMed ii) enhances response accuracy using RAG approaches. Through natural language queries, users can retrieve information about viruses, their hosts and ecosystems and other associated metadata (e.g., matrix, location, sample collection), helping researchers navigate complex datasets without requiring advanced query languages or specialized bioinformatics expertise. Strict safeguards minimize LLM hallucinations and secure sensitive data, including bioterrorism-related risks.

The resource is publicly available and running on a biosphere cloud instance hosted by IFB/PRABI. The demo will cover i) the tool's architecture, ii) integration of external tools with LLMs, and iii) its connection to a Neo4j graph database, with use cases showcasing conversational AI for viral data exploration.

URL

https://prabi-cloud149.univ-lyon1.fr/ or https://viromeatlas.univ-lyon1.fr/ai

Password : Trojan_horse37866

Posters

{affiliationExplorer} a Shiny webapp to resolve taxonomy conflicts

Poster

Mahendra Mariadassou [1], ***Sandra Dérozier*** [1], ***Cédric Midoux*** [1], ***Olivier Rué*** [1]

1. *INRAE*

Abstract

Metabarcoding studies in microbial ecology generate taxonomic abundance profiles, with each taxon usually represented by an ASV (*Amplicon Sequence Variant*). Some taxa may be associated with several, sometimes conflicting, taxonomic affiliations. These ambiguities (or multi-affiliations), may be frequent when analysing short sequences and/or when reference databases are incomplete. This complicates ecological interpretation and decreases the reproducibility of results. According to the origin of the samples (*e.g.* food, fecal, biowaste, etc.) or the conflict source (spelling typo in the reference databases) , manual curation can sometimes resolve these ambiguities. Likewise, some inconsistencies can be easily corrected programmatically if a manual curation was performed on a similar dataset and needs to be reapplied. The ‘{affiliationExplorer}‘ R-package provides a user interface dedicated to exploring and resolving these conflicts, and can easily be inserting in BIOM (*Biological Observation Matrix*) data analysis workflows.

Developed in Shiny with the {golem} framework, this package offers a user-friendly interface for ① loading a BIOM file, a multi-affiliation file, and, optionally, one or more expert-verified sequence–taxonomy dictionaries (in FASTA format), ② visualising and ordering multi-affiliated taxa by abundance, ③ automatically resolving conflicts using the provided dictionaries for curated ASV references, ④ performing expert manual curation by selecting the most relevant affiliation among those proposed or specifying it manually, and ⑤ exporting a corrected BIOM file and a sequence–taxonomy dictionary to log the performed curations and for potential reuse in subsequent curation sessions. The produced BIOM file follows the standards of biom-format.org and is fully compatible with standard metabarcoding workflows. The exported FASTA files can be reused as is in the interface.

URL

https://shiny.migale.inrae.fr/app/affiliationexplorer

A benchmark dataset for analyzing the functional fate of duplicate gene pairs in the model plant Arabidopsis thaliana

Poster

Erine Benoist [1], ***Samuel Ortion*** [2], ***Séanna Charles*** [3], ***Emmanuelle Lerat*** [4], ***Franck Samson*** [5], ***Marie Szafranski*** [6], ***Carène Rizzon*** [1]

1. *Université d'Evry Paris-Saclay,* ***2.*** *Genoscope, Institut François Jacob, CEA, CNRS, Univ Evry, Université Paris-Saclay, Evry, France,* ***3.*** *Université Claude Bernard Lyon 1,* ***4.*** *CNRS,* ***5.*** *INRAE,* ***6.*** *ENSIIE*

Abstract

The prediction of the functional fate of duplicated genes in *Arabidopsis thaliana* has recently been addressed using machine learning approaches (Ezoe et al., MBE 2020; Meinke, New Phytologist 2020). However, improving the performance and generability of such approaches requires datasets that represent more comprehensively the full repertoire of duplicate genes. To address this limitation, we expanded the dataset originally compiled by Ezoe et al. (MBE 2020), which contained labeled gene pairs with binary labels (0/1) indicating their functional fate, by incorporating duplicate gene pairs identified across the entire genome, thereby generating a more representative dataset for predictive modeling.

To achieve this, we implemented a pipeline for the identification of duplicated genes, complemented by a merging step that enables the definition of supplementary gene pairs. In parallel, the labeled dataset was expanded from the initial dataset published by Ezoe et al. (MBE 2020) by incorporating additional annotations from another study (Meinke, New Phytologist 2020; Bolle et al., The Plant Journal 2013). This approach resulted in the establishment of a curated benchmark dataset providing a robust reference for the definition and analysis of 75,631 duplicated gene pairs.

In addition, several descriptors were updated relative to the original study, and new features were computed to enable the use of this benchmark dataset for predictive modeling. These descriptors include the synonymous and non synonymous substitution rates (Ks and Ka), gene expression profiles, PFAM domain annotations, transposable element (TE) density and coverage, and protein–protein interaction (PPI) data.

A bioinformatics pipeline for de novo detection of tandem repeats in common bean genomes

Poster

Maisen Hassani [1], ***Valerie Geffroy*** [1], ***Gianluca Teano*** [1]

1. *INRAE*

Abstract

The advent of long-read sequencing technologies, particularly those developed by Oxford Nanopore Technologies (ONT) and Pacific Biosciences (PacBio HiFi), has greatly improved plant genomics by enabling more complete and continuous genome assemblies, including highly repetitive regions. The common bean, Phaseolus vulgaris, the most important grain legume for human consumption now benefits from several high-quality genome assemblies: 12 assemblies generated using ultra-long ONT reads and 15 assemblies obtained using PacBio HiFi reads.

In this context, we developed a bioinformatics pipeline to identify de novo tandem repeats in these genome assemblies. Several repeat detection tools were compared (TRF, ULTRA, TideHunter and TRASH2) using the reference genome assembly from BAT93. Among them, TRF was selected as the best tool because it detected the largest number of tandem repeats.

The TRF results were then used to study the distribution of repeats along the chromosomes. Peak detection methods were applied to identify genomic regions with a high density of repeats, highlighting regions of potential interest.. As a validation step, the detected motifs were compared to known centromeric satellites CentPv1 and CentPv2 of the common bean genome using BLAT alignments. Hits with at least 90% coverage and 88% accuracy were kept for further analysis.

To confirm the presence and organization of repeats in the selected regions, similarity triangles were generated to visualize the local repeat structure. Finally, a manual inspection of the results allowed the identification of genomic windows of interest. The selected regions were then used to extract the different repeated sequences detected by TRF.

After stabilizing the pipeline and optimizing its parameters, it was used to identify repeats in other common bean genomes from different origins (equatorial, northern and southern). This approach allows the exploration of tandem repeat diversity and organization across common bean genomes and the identification of potential new repetitive elements.

URL

1) https://forge.inrae.fr/gedy/foreverbeanmaisen

2) https://github.com/Benson-Genomics-Lab/TRF

3) https://academic.oup.com/nar/article/27/2/573/1061099

A Course-Undergraduate Research Experience (CURE) to explore the effect of structural variants on gene expression in C. elegans balancers

Poster

Tatiana Maroilley [1], ***Victoria Rodrigues Alves Barbosa*** [2], ***Rumika Mascarenhas*** [2], ***Suzanne Ferris*** [2], ***Catherine Diao*** [2], ***Consortium Students MDSC 301 2023*** [2], ***David Anderson*** [2], ***Maja Tarailo-Graovac*** [2]

1. *IGBMC,* ***2.*** *University of Calgary*

Abstract

Bioinformatics, a discipline at the crossroads of Biology and Computational Sciences, also referred to as Computational Biology, is nowadays widely spread in research programs. However, implementing any Bioinformatics projects requires the ability to comprehend biological concepts and apply computational approaches, and rare are the undergraduate programs offering such multi-disciplinary training. In addition, understanding the dynamic between Biology research projects and Bioinformatics analyses is challenging with no real-life experience. Course-based undergraduate research experience (CURE) courses are innovative programs that allow more students to acquire research experience and provide the perfect setting to introduce students to applied bioinformatics. As a part of the Bachelor of Health Sciences of the Cumming School of Medicine at the University of Calgary (Canada), a CURE applied bioinformatics was implemented in the Winter of 2023 to 2025. Students investigated the effect of structural variants (SVs, genetic variants larger than 50 bp) on gene expression in the model organism *Caenorhabditis elegans* (a hermaphrodite 1-mm long roundworm). The students detected and characterized SVs by analyzing genome and transcriptome sequencing data of *C. elegans* strains called balancers, as they are known to carry large genomic variations balancing regions of the genome by limiting recombination and allowing maintenance of lethal mutations. They used Galaxy, a public web-based supercomputing resource, but also a local High-Performance computing system, and R, to report different effects of SVs on gene expression and splicing. Students' research explained the molecular mechanism behind the uncoordinated phenotype caused by the reciprocal translocation eT1(III;V) and uncovered unexpected effects on gene expression of an understudied gene. We evaluated the course's impact on student learning journeys and showed that the CURE favored students' understanding of the Bioinformatics field and fostered their research interest. We provide guidelines here to facilitate the CURE implementations to improve access for undergraduate students to bioinformatics research experiences.

URL

https://www.biorxiv.org/content/10.64898/2026.01.21.700799v1

A long-read metagenomic pipeline for deciphering yam virome: overcoming host-integrated sequences challenges

Poster

Maimouna Kone [1], ***Estel Pakyendou NAME*** [2], ***Ezechiel TIBIRI*** [2], ***Fidele Tiendrebeogo*** [3], ***Justin S. PITA*** [4]

1. *Laboratoire d'Innovation pour la Santé des Plantes, UFR Biosciences, Université Félix Houphouët Boigny, Abidjan,* ***2.*** *Laboratory of Virology and Plant Biotechnology, Institute of Environment and Agricultural Research (INERA), 01BP476,* ***3.*** *Central and West African Virus Epidemiology (WAVE), Pôle scientifique et d'innovation de Bingerville, Université Félix Houphouët-Boigny (UFHB),* ***4.*** *Central and West African Virus Epidemiology (WAVE), Pôle scientifique et d'innovation de Bingerville, Université Félix Houphouët-Boigny (UFHB), Bingerville, Côte d'Ivoire*

Abstract

Yams (*Dioscorea spp*.) are vital food crops in West Africa particularly in Côte d'Ivoire, yet their production is severely threatened by Badnaviruses. A major computational challenge in their detection arises from the presence of endogenous pararetroviral sequences (EPRV) integrated into the host genome, which often confound traditional short-read diagnostic methods and lead to false positives [1].

We developed a dedicated bioinformatic pipeline (https://github.com/etibiri/denovo-assembly-pipeline) designed to selectively characterize episomal viral genomes from long-read metagenomic data. To enrich circular viral DNA, Rolling Circle Amplification (RCA) was performed on 20 samples from Côte d'Ivoire, followed by Oxford Nanopore Technology MinION sequencing. The computational framework integrates:

- Raw read processing and filtering: Quality control and length-based filtering to optimize assembly.
- *De novo* viral assembly: Leveraging long-read capabilities to bridge complex repetitive regions common in Badnaviruses.
- Genomic differentiation: A comparative sequence analysis strategy to distinguish episomal viral contigs from endogenous host sequences.
- Taxonomic and Phylogenetic Inference: Automated identity clusters and maximum-likelihood phylogenetic reconstruction.

The pipeline successfully reconstructed three complete episomal genomes (7.3–7.5 kbp) and identified 12 high-quality partial sequences. Sequence identity analyses revealed 91% similarity with Dioscorea alata bacilliform virus and 86-89% with Dioscorea bacilliform RT virus. Phylogenetic placement categorized these isolates into four distinct groups (T15, K8, K9, and K5) [2].

This workflow demonstrates high sensitivity in detecting mixed-viral infections and provides a reproducible, scalable tool for pathogen surveillance. By overcoming the limitations of short-read technologies and host-sequence interference, this bioinformatic approach provides a robust foundation for real-time monitoring of viral diversity in complex plant genomes.

URL

https://www.mdpi.com/1999-4915/17/12/1586

A mathematical framework to accurately reconstruct cell lineage from single cell transcriptomics on barcoded cells: application for therapeutics optimization

Poster

Anne-Sophie Giacobbi [1], ***Bence Hadju*** [1], ***Annabelle Ballesta*** [1]

1. *Institut Curie*

Abstract

Introduction: Quantifying tumor plasticity is essential for understanding resistance in glioblastoma (GBM), the most frequent and aggressive brain tumors in adults. A mathematical pipeline is presented for the joint analysis of time-resolved single-cell (sc) transcriptomics and lineage barcoding to characterize cancer cell states and quantify their dynamics (proliferation, death, state transitions) in control or temozolomide-treated conditions. This model may further be used to design interventions targeting specific cell states to maximize drug efficacy.

Methods: Cell states are identified through clustering and enrichment analysis of known signatures. An ordinary differential equation (ODE)-based model is utilized to infer cell state dynamics by integrating both sc barcoding and transcriptomics. Existing work [1] was extended through the incorporation of exponential growth, improved parameter optimization via the CMA-ES algorithm and the relaxation of sparsity constraints.

Results: In GBM cell lines, lineages are reconstructed across three time points into a hierarchical tree that achieved a close fit to data. Estimating parameters in control or temozolomide-treated conditions revealed key resistance mechanisms. The pipeline was benchmarked against the original model. Strategies to maximize efficacy were designed to predict optimal interventions on cell state death or transition rates. Identifying corresponding molecular targets remains the next challenge to allow for clinical translation of these findings.

Reference:

[1] Larsson et al., Modeling glioblastoma heterogeneity as a dynamic network of cell states, Molecular systems biology, 17 (9), 2021

A Multiomic Atlas of Human Microprotein-Coding Intronic Polyadenylation Isoforms

Poster

Matthaus Sirvent [1], ***Nicolas Fontrodona*** [1], ***Celine Labbe*** [2], ***Didier Auboeuf*** [1], ***Martin Dutertre*** [2]

1. *Laboratoire de Biologie et Modélisation de la Cellule, École Normale Supérieure de Lyon, CNRS, UMR 5239, Inserm, U1293, Université Claude Bernard Lyon 1,* ***2.*** *Institut Curie, PSL Research University, CNRS UMR 3348, INSERM U1368*

Abstract

Background

Recent advances reveal that the human proteome is more complex than previously thought. Regions annotated as noncoding (*e.g.*, noncoding RNAs and untranslated regions [UTRs]) contain open reading frames (ORFs) whose translation is supported by Ribo-seq and mass spectrometry. Most of these non-annotated ORFs encode microproteins (miPs) of about 100 amino acids or less, forming a “dark proteome” whose size remains intensely debated.

Intronic Polyadenylation (IPA) generates transcript isoforms ending upstream of the last exon of genes. We recently reported that IPA isoforms terminating within an annotated 5’UTR intron are translated into miPs. Currently, no resource exists for these microprotein-coding, 5’UTR-located intronic polyadenylation (miP-5’UTR-IPA) isoforms.

Results

We present miP-IPA DB, the first database of miP-5’UTR-IPA isoforms. This multiomic atlas focuses on **Homo sapiens**.

Each miP-IPA entry of the database is located between the 5’-most transcription start site and the annotated coding sequence (CDS) of a gene. This ensures that the identified miPs are unique products of annotated non-coding regions rather than truncated isoforms of known proteins. The datatabase integrates IPA isoforms supported by 3’seq and long-read RNA-seq data with ORFs supported by Ribo-seq and/or mass spectrometry data, to ensure multievidence support.

We identified hundreds of miP-5’UTR-IPA isoforms and encoded miPs. These transcript isoforms are often expressed in multiple tissues and cell types, both normal and cancerous, and regulated by anticancer drugs. Significant subsets of ORFs are conserved in chimp, present homology with other human proteins and/or contain predicted functional motifs.

Conclusion

miP-IPA DB serves as a resource for the study of human miP-5’UTR-IPA isoforms and their contribution to the dark proteome, by formalizing the relationship between alternative polyadenylation and microprotein translation. In addition, our data suggest potential roles of miP-5’UTR-IPA isoforms in physiology and cancer.

A reproducible genomic and predictive modelling framework for characterising clinical antimicrobial resistance: A long-read sequencing study in Burkina Faso

Poster

Nènè Sthella KY [1], ***Ezechiel TIBIRI*** [1], ***Estel Pakyendou NAME*** [1], ***Marguerite Edith Malatala NIKIEMA*** [1], ***Pamane DJAGBARE*** [2], ***Wendyam Marie Christelle NADEMBEGA*** [2], ***Lassina TRAORE*** [2], ***Emmanuel SAMPO*** [3], ***Moussa OUEDRAOGO*** [3], ***Fidele Tiendrebeogo*** [4], ***Jacques SIMPORE*** [2]

1. *Laboratory of Virology and Plant Biotechnology, Institute of Environment and Agricultural Research (INERA), 01BP476,* ***2.*** *Laboratory of Molecular Biology and Genetics, LABIOGENE, University Joseph KI-ZERBO,* ***3.*** *Schiphra Hospital Biomedical Laboratory,* ***4.*** *Central and West African Virus Epidemiology (WAVE), Pôle scientifique et d'innovation de Bingerville, Université Félix Houphouët-Boigny (UFHB)*

Abstract

Antimicrobial resistance is a major public health challenge, particularly in low-resource settings where genomic surveillance remains limited. In Burkina Faso, urinary tract and suppurative infections are frequent, with *Escherichia coli* and other Enterobacteriaceae among the predominant pathogens. This study aimed to develop a reproducible genomic and predictive modelling framework for characterising clinical antimicrobial resistance using long-read sequencing data.

Thirty-nine clinical isolates from urine and pus samples collected at HOSCO and SCHIPHRA hospitals were sequenced using Oxford Nanopore Technologies. The workflow integrated automated quality control and statistics using SeqKit and NanoPlot, de novo assembly with Flye, iterative Medaka polishing assessed by BUSCO, taxonomic assignment with Diamond and Bakta gene prediction, and resistome profiling with ResFinder using stringent thresholds. Plasmid characterisation was performed using MobSuite. The complete workflow was implemented in a reproducible framework (https://github.com/kysthella/CIBIG_2025).

Sequencing quality strongly influenced downstream analyses. High-yield barcodes, mainly from the 20–39 series, generated complete or near-complete genomes, whereas low-yield samples produced fragmented or failed assemblies. Six high-quality complete *E. coli* genomes and one *Enterobacter cloacae* complex genome were retained for comparative analyses. Resistome analysis revealed multidrug resistance profiles among *E. coli* isolates, including beta-lactamase-associated determinants such as *bla*CTX-M genes and markers affecting several antibiotic classes. MLST identified diverse *E. coli* sequence types, including ST2659, ST38, ST44 and ST2014. MobSuite analysis showed a predominance of IncF-type plasmid replicons and frequent conjugative mobility, supporting the relevance of plasmid characterisation in AMR surveillance.

Finally, AMR and plasmid profiles were integrated into exploratory downstream analyses, including resistome clustering, calibrated simulation and MDR-like classification. This reproducible framework supports genomic AMR surveillance and future predictive modelling in resource-limited settings.

URL

https://github.com/kysthella/CIBIG_2025

A snakemake pipeline to genotype large sets of short reads on a pangenome using pangenie

Poster

Martin RACOUPEAU [1], ***Frederic CHOULET*** [2], ***Fabrice Legeai*** [3], ***Christine Gaspin*** [4], ***Christophe Klopp*** [5]

1. *Université de Toulouse, INRAE, UR 875 MIAT, F-31320, Castanet-Tolosan, France,* ***2.*** *UMR 1095 GDEC, INRAE ; Université Clermont Auvergne,* ***3.*** *UMR 1349 IGEPP, INRAE, Le Rheu 35650, France ; Univ Rennes, Inria, CNRS, IRISA - UMR 6074, F-35000 Rennes, France,* ***4.*** *Université de Toulouse, INRAE, UR 875 MIAT, F-31320,* ***5.*** *Sigenae, MIAT UR875, INRAE, F-31326, Castanet Tolosan, France.*

Abstract

Pangenome graphs improve variant detection in regions not or underrepresented in a linear reference assembly and enable the identification of structural and complex variants. Once identified, these variants can be genotyped across large datasets using short reads [1, 2]. Numerous short-read data sets are publicly available for many species of interest.

Recent mapping- and k-mer–based tools enable efficient genotyping from individual datasets. However, scaling these analyses to large collections datasets and efficiently merging genome-wide VCFs remains computationally challenging.

To address this, we developed a Snakemake workflow to automate variant indexing and genotyping. The workflow supports both local datasets and lists of public SRA accessions, automatically retrieving the corresponding sequencing files. It dynamically manages disk space and memory usage, enabling the processing of large datasets in constrained computing environments. The workflow is fully containerized using Apptainer and supports parallel execution.

The current implementation integrates PanGenie [4] and processes VCF files generated with Pan1C [6] or Minigraph-Cactus [7].

The workflow outputs VCF files containing merged genotypes per chromosome. It allows large-scale genotyping while maintaining reproducible and resource-aware execution.

[1] Sirén et al. 2021. *Science*. doi:10.1126/science.abg8871

[2] Liao et al. 2023. *Nature*. doi:10.1038/s41586-023-05896-x

[3] Garrison et al. 2018. *Nat. Biotechnol.* doi:10.1038/nbt.4227

[4] Ebler et al. 2022. *Nat. Genet*. doi:10.1038/s41588-022-01043-w

[5] Du et al. 2025. *Mol. Plant.* doi:10.1016/j.molp.2025.08.001

[6] Mergez et al. 2025. *HAL*. hal:05034842

[7] Hickey et al. 2023. *Nat. Biotechnol.* doi:10.1038/s41587-023-01793-w

URL

https://forge.inrae.fr/martin.racoupeau/pangenome_genotyping/-/tree/main

AI-stro: a neuro-symbolic approach to astrocyte regulation in artificial neural networks

Poster

Nathan Olejniczak [1], ***Arnaud Kress*** [1], ***Luc Moulinier*** [1], ***Alexandre Charlet*** [1], ***Anne Jeannin-Girardon*** [1], ***Hugues Petitjean*** [1]

1. *Université de Strasbourg*

Abstract

Astrocytes are glial cells that actively regulate neuronal activity through local and spatially organized mechanisms, controlling excitability, modulating firing dynamics, and operating over groups of neurons rather than individual cells. Despite their central role in biological neural activity, astrocytic regulation has no counterpart in standard machine learning architectures. This paper presents a prospective neuro-symbolic framework that formalizes astrocytic mechanisms as symbolic constraints on the geometry of an artificial neural network.

The core proposal is to govern the slope of the positive piece of the rectified linear activation function (ReLU) through explicit, biologically-grounded rules. Unlike parametric approaches that learn activation shapes purely to optimize task performance, our framework subordinates architectural learning to domain knowledge: permissible activation geometries are determined by what the biology allows, at the level of individual neurons and neuronal groups.

Validation is twofold: the constrained architecture must not impair learning on standard tasks, and its activity dynamics must be comparable to neural activity time series recorded in vivo. The conjunction of these criteria, functional and dynamic, is what makes the framework both scientifically tractable and practically meaningful.

A longer-term horizon concerns energy efficiency: if astrocytic regulation contributes to the brain's metabolic parsimony, biologically-grounded constraints may point toward artificial systems that are more interpretable and less computationally demanding.

AlphaFold-Multimer Predictions : Which Scores Best Identify True Protein–Protein Interactions in the TGF-β Activation Network?

Poster

Elisa Chenel [1], ***François COSTE*** [1], ***Samuel BLANQUART*** [1], ***Catherine BELLEANNÉE*** [1], ***Nathalie Théret*** [2]

1. *Univ Rennes, Inria, CNRS, IRISA - UMR 6074, Rennes, F-35000, France,* ***2.*** *Univ Rennes, INSERM, EHESP, IRSET - UMR 1085, Rennes, F-35000, France ; Univ Rennes, Inria, CNRS, IRISA - UMR 6074, Rennes, F-35000, France*

Abstract

Protein-protein interactions (PPIs) play a central role in cellular processes. Changes in PPIs such as mutations or the absence of a partner can lead to cellular dysfunction and diseases. Identifying these interactions and obtaining their structures is therefore crucial for the development of therapeutic strategies.

Recent advances in computational methods, notably AlphaFold and, by extension, AlphaFold-Multimer, have enabled large-scale prediction of protein complexes from sequence information.

The predicted structure corresponds to the most plausible structure based on the sequences, however it does not ensure that the predicted structure is a true interaction.

The state-of-the-art approach consists in applying thresholds to the scores associated with structure predictions to distinguish the likely interactions from those that are not.

In this study, we evaluated whether if some of these scores can be used to distinguish experimentally validated PPIs from other predicted interactions, using our expertise on the PPI network involved in the activation of the profibrotic factor, TGF-β.

Result

The evaluation of the scores using ROC curve showed their limitation in distinguishing validated PPIs (ROC curve's AUC close to 0.5) Therefore, none of the scores outperforms the others in terms of AUC. However, the precision-recall curves showed a moderate enrichment in valid PPIs among the predictions with the best scores. Hierarchical clustering of scores revealed the existence of a prediction subgroup with high scores involving related proteins.

These results highlight the limitations of current confidence scores for globally discriminating PPIs in the TGF-β activation network. Interpreting the scores of a predicted structure based on key biological features of the involved proteins (size, disordered regions…) may improve the identification of true PPIs.

An integrated long-read bioinformatics pipeline for resolving the genetic diversity of Plasmodium falciparum csp in Burkina Faso

Poster

Emilie S BADOUM [1], ***Ludovic KOURAOGO*** [1], ***Jean W SAWADOGO*** [1], ***Issa Nebié OUEDRAOGO*** [1], ***Alfred B TIONO*** [1], ***Alphonse OUEDRAOGO*** [1], ***Sodiomon B SIRIMA*** [1]

1. *Groupe de Recherche Action en Santé*

Abstract

Malaria remains a major public health concern in Burkina Faso where intense seasonal transmission may shape the genetic landscape of *Plasmodium falciparum*. The circumsporozoite protein (csp), a key pre-erythrocytic antigen and the main target of the RTS, S/AS01 and R21 vaccines exhibits extensive polymorphism that may influence vaccine performance and parasite population dynamics [1,2].

Csp amplicons were sequenced using ONT R10 flow cells and processed using the NanoRave platform. High accuracy basecalling (Q>10) and stringent QC preceded consensus generation. Variants (SNPs/indels) were identified using an ONT optimized workflow combining Minimap2 and Medaka/Longshot, followed by domain resolved haplotype reconstruction across the N-terminal, central repeat and C-terminal regions. Unlike standard pipelines, our approach emphasizes resolving csp repeat motifs enabling more accurate estimates of nucleotide diversity (π) and Watterson's theta (Θ_w) through full-length haplotype recovery in line with ONT amplicon studies [3,4].

Analysis of Banfora samples collecting during dry and rainy seasons revealed 27 high-confidence polymorphisms. Computational metrics indicated higher genetic complexity during the peak transmission season (Θ_w: 5.22 vs. 4.72). A strongly negative Tajima's D (-2.95) was consistently detected, which suggests either recent population expansion or purifying selection. Our pipeline successfully distinguished between seasonal fluctuations and stable polymorphic sites, providing a high-resolution map of variability in vaccine targets.

This study shows that using long-read sequencing with a specialized bioinformatics pipeline offers a scalable, reproducible solution for real-time genomic surveillance of malaria. This framework can be easily adapted to other highly polymorphic vaccine candidates including in resource limited settings.

An integrated R package for interpretable deep learning on multi-omics data in system immunology.

Poster

Philippe STOCKER [1], ***Nicolas Tchitchek*** [1]

1. *INSERM UMRS 959, Immunology-Immunopathology-Immunotherapy (i3), Sorbonne Université, Paris, France*

Abstract

Over the last decades, large-scale multi-omics datasets have been generated across many fields of the life sciences and medical biology. The challenge with such data is leveraging them to identify biomarkers that discriminate patient subgroups, predict clinical responses to therapies, and integrate heterogeneous omics layers into unified analytical frameworks. Neural networks offer a powerful approach to address these challenges by learning representations that capture complex biological patterns. However, implementing robust and interpretable deep learning workflows remains challenging.

We developed an R package that streamlines the deployment of deep learning workflows for multi-omics datasets. The framework is built on an object-oriented architecture using R6 and wraps the Keras 3 API backed by TensorFlow, providing a robust modeling engine within the R ecosystem. This R package supports the implementation of complete deep learning pipelines, including classification and regression models that link biological features to clinical labels for disease classification and treatment response prediction, as well as autoencoders that compress data to learn representations within a latent space. The framework handles k-NN imputation for sparse datasets. Additionally, it uses feature selection using Cohen's d or Cliff's Delta to prioritize the most relevant biological signals. The package also integrates visualization methods, including dimensionality reduction techniques for latent space exploration. Furthermore, SHAP-based methods have been implemented to allow results interpretability through the fastshap and shapviz packages.

The package is designed as an accessible and flexible wrapper around deep learning workflows for applications to a wide range of biological and biomedical studies. Model architectures and hyperparameters can be customized at will by the user. The framework can automatically adapt model structures according to the characteristics of the input dataset. The framework allows researchers to perform predictive modeling, explore latent representations of biological data, and interpret model outputs within a unified workflow.

URL

https://github.com/tchitchek-lab/ImmunoDeepR

An Integrative Deep Learning and Structural Workflow for Accurate Annotation of Insect Odorant Receptors

Poster

David Gilardot [1], ***Audrey Chathuant*** [1], ***Camille Meslin*** [2], ***Nicolas Montagné*** [3], ***Emmanuelle Jacquin-Joly*** [1]

1. *Institut National de Recherche pour l'Agriculture, l'Alimentation et l'Environnement ; Institut d'Ecologie et des Sciences de l'Environnement de Paris, Versailles cedex 78026, France,* ***2.*** *Institut National de Recherche pour l'Agriculture, l'Alimentation et l'Environnement ; d'Ecologie et des Sciences de l'Environnement de Paris, Versailles cedex 78026, France,* ***3.*** *Institut National de Recherche pour l'Agriculture, l'Alimentation et l'Environnement ; Sorbonne Université ; Institut d'Ecologie et des Sciences de l'Environnement de Paris ; Institut Universitaire de France*

Abstract

In insects, olfaction is vital for many behaviors, including host search, enemy avoidance, and reproduction. Odorant receptors (ORs) are key proteins in this process, and the OR gene superfamily is evolving fast in a birth-and-death model, with a resulting average sequence identity of only 30%[1]. This precludes their correct annotation by current automatic pipelines in omics data, and manual annotations still remain the standard procedure.

Here, we present the development of a novel integrative bioinformatic workflow based on ab initio and deep learning (DL) methods that exploits the sequences and the three-dimensional structures of ORs to correctly annotate them in insects.

Applied on the moth Spodoptera frugiperda, the DL model - trained specifically on Lepidoptera ORs - enabled the identification of nearly 85% out of the 71 OR genes manually identified previously[2]. Combined with ab initio approaches, we achieved 95% completeness. Applied on the aphid Myzus persicae, our pipeline discovered two news ORs, detected six pseudogenes, and improved the annotation of six ORs compared to previous manual annotations[3].

Our workflow provides a potentially generic method for insect OR annotation, an essential step for future studies on OR functioning and evolution. From an applied point of view, this workflow opens the way for in silico molecular docking and OR ligand prediction[4], enabling the identification of semiochemicals to be used in environmental-friendly control strategies.

Are deep learning methods accurate to predict protein functions in marine organisms?

Poster

Rodrigo Salinas [1], ***Perrine Kergoat*** [1], ***Laurence Garczarek*** [1], ***Frederic Partensky*** [1], ***Fabio Rocha Jimenez Vieira*** [1], ***Juliana Bernardes*** [1]

1. *Sorbonne Université, CNRS, UMR 7144, Adaptation & Diversity in the Marine Environment, Station Biologique de Roscoff (SBR), 29680, Roscoff, France.*

Abstract

Proteins play essential roles in living organisms, catalyzing biochemical reactions, transmitting signals, and maintaining cellular structure. Although protein function can be determined experimentally, the cost and effort of these approaches relent the process of function discovery. Several deep learning (DL) models, particularly those based on protein language models, have recently been proposed and showed improvements over traditional approaches. However, most DL models have been evaluated on benchmarks of highly divergent proteins and are rarely compared with established tools such as InterProScan, a highly precise method based on profile hidden Markov models.

Here, we evaluated five DL-based tools (DeepGOPlus, ProteInfer, FANTASIA, PLMSearch, and ProteNN2) using two datasets: Original ***(OG)***, containing 4,380 poorly annotated proteins from alpha-cyanobacteria reference proteomes, and Shuffled ***(SH)***, an amino acid shuffled-version of ***OG*** sequences as negative control. Performance was assessed by each method's ability to distinguish real from shuffled sequences. We defined the False Discovery Rate (**FDR**) as the ratio of predictions on ***SH*** to the total number of proteins in ***OG*** and computed prediction coverage and FDR across sorted confidence scores and Gene Ontology tree depths.

Our results show that the performance of the evaluated methods depends on the proportion of false predictions observed. At FDR below 10%, InterProScan outperformed most DL tools and achieved a coverage of 55,9% of proteins in ***OG***, only tied with DeepGOPlus. At FDR=15%, PLMSearch outperformed InterProScan, with a protein coverage of 60,6%. Finally, at FDR=20%, ProteInfer slightly surpassed InterProScan with a 56,7% coverage, FANTASIA and ProteNN2 remained below these coverage even at FDR=20%. Overall, DL tools produced more false positives and more generic GO terms than InterProScan at FDR <10%. These findings highlight the need for more accurate DL approaches with lower FDR while helping users select protein function prediction tools according to their desired balance between coverage and precision.

ArmVar: a novel approach to identify cancer cells from single-cell RNA-sequencing datasets

Poster

Mehdi Marchand [1], ***Lucie Lamothe*** [2], ***Yuna Blum*** [3], ***Remy Nicolle*** [1]

1. *Université Paris Cité, Centre de Recherche sur l'Inflammation (CRI), INSERM, U1149, CNRS, ERL 8252, F-75018 Paris France,* ***2.*** *Laboratoire D'informatique de Grenoble,* ***3.*** *IGDR*

Abstract

Since tumors are composed of a mixture of normal and transformed cells, the analysis of tumors using single-cell RNA sequencing (scRNA-seq) data first requires to accurately identify transformed cancer cells. Different strategies were proposed to identify malignant cells using specific cancer cell markers combined with cell partitioning. However, cancer markers are often cancer-type dependent and markers of subsets of their normal cellular counterparts, making these methods limited.

On the other hand, other approaches aim to identify cancer cells based on variation of gene expression reflecting Copy Number Variations (CNVs) (e.g. FastCNV, CopyKAT). These genomic alterations correspond to gain or loss of genomic regions which are associated with diseases such as cancer. In general, methods inferring CNVs based on scRNA-seq data assume that higher or lower expression compared to normal conditions may reflect the presence of additional copies or the absence of genomic regions, respectively. Nevertheless, these associations are quite indirect since gene expression is regulated by complex transcriptional mechanisms.

If some methods can infer CNVs without reference, the use of a reference seems to increase the performance. Thus, methods can be divided into two groups, namely those that require a reference to be specified by the user (supervised), and those determining the reference internally (unsupervised).

In this study, we introduce ArmVar, a novel unsupervised approach based on a chromosomal arm-level mean expression test using Welch's ANOVA, in combination with smoothing based on the annotations of neighboring cells determined through nearest neighbor graph construction. We compare this method with two state-of-the-art approaches, fastCNV and CopyKAT. Applied to four annotated scRNA-seq datasets from the literature, our method shows particularly encouraging results for the identification of cancer cells. It achieves better results than other existing unsupervised methods and comparable results to those of the tested supervised approach.

Assessing Dorado pseudourydilation RNA modification prediction on Arabidopsis thaliana ribosomal RNA

Poster

Emma Rodriguez [1], ***Anne de Bures*** [2], ***Adrien Castinel*** [3], ***Benjamin Charlier*** [1], ***Virginie Marchand*** [4], ***Yuri Motorin*** [4], ***Céline Vandecasteele*** [3], ***Nathalie Vialaneix*** [1], ***Julio Sáez Vásquez*** [2], ***Christine Gaspin*** [1]

1. *Université de Toulouse, INRAE, UR 875 MIAT, F-31320,* ***2.*** *LGDP, Université de Perpignan Via Domitia, F-66860,* ***3.*** *GeT-PlaGe, Genotoul, INRAE, US1426,* ***4.*** *SMP_IBSLor, Université de Lorraine, F-54505*

Abstract

Epitranscriptomics refers to the collection of chemical modifications that affect RNA without altering its sequence. To date, more than 170 RNA modifications have been described and growing evidence highlights their key role in a wide range of biological processes [1]. Direct RNA sequencing from Oxford Nanopore Technologies (ONT) is a recent technology enabling the production of long and reliable reads. It is particularly promising because it can, in principle, detect any RNA modifications across the entire epitranscriptome. However, several methodological challenges still need to be addressed to fully exploit this technology. To convert raw electrical signals into annotated RNA bases, the Dorado software performs both basecalling and RNA modification prediction with deep-learning models released by ONT. However, its performance depends on factors such as signal noise, sequence context, and the diversity of organisms and datasets used for training [2]. Previous evaluations have shown that detection models may generate false positives [3] and that robust large-scale models are currently available for only a limited number of RNA modifications [4].

We generated a four-sample *Arabidopsis thaliana* direct RNA sequencing dataset and compared Dorado predictions with a literature-derived dataset describing 105 pseudouridylation (Ψ) sites in ribosomal RNAs (18S, 5.8S, and 25S rRNAs) [5,6,7,8]. We then computed the distribution of proportions of reads predicted as modified at positions containing a uridine (U).

Preliminary results from Figure 1 illustrate the extent to which the sites reported in the literature (blue) are recovered. Using a 5% Type-I error threshold on the proportion of reads predicted as modified by Dorado, the proportions of common, dorado-specific and literature-specific Ψ sites are respectively 51.9% (83 sites), 34.4% (55) and 13.8% (22). We plan to extend this work to a broader scale analysis, including additional RNA modifications, experimentally validating predicted modification sites and improving filtering strategies to reduce false-positive rate.

Assessing the structure of DNA embedding spaces using graph-based comparisons

Poster

Juliette Francis [1], ***Quentin Le Graverand*** [2], ***Mahendra Mariadassou*** [3], ***Yann Le Cunff*** [1]

1. *Univ Rennes, Inria, CNRS, IRISA - UMR 6074, Rennes, F-35000, France,* ***2.*** *GenPhySE, Université de Toulouse, INRAE, ENVT, 31326, Castanet Tolosan, France,* ***3.*** *Université Paris-Saclay, INRAE, MaIAGE, 78350, Jouy-en-Josas, France*

Abstract

Many models have been proposed to create embeddings of DNA sequences. While these models are typically evaluated using downstream tasks such as species prediction, such evaluations offer limited insight into the intrinsic differences between their embedding spaces. To address this gap, we focus on direct comparison of the geometric and topological properties of embeddings generated by those models.

We consider five models (Bag-of-kmers, dna2vec, DNABERT-S, DNABERT2 and HyenaDNA) chosen to to represent a broad range of embedding techniques: from simple k-mer counts to state-of-the-art transformer architectures. Our comparison centers on two key questions. First, how do these models organize sequences from the same species in their latent spaces ? Second, how do they differ in terms of local and global topological properties ?

We constructed k-nearest neighbors (K-NN) graphs and applied two complementary analyses: (i) Jaccard distance to quantify local neighborhood similarities and (ii) a permutation test based on Random Dot Product Graphs (RDPG) to compare global topological properties.

In summary, our study establishes Jaccard distance and RDPG on k-NN graphs as a unified framework for comparing DNA sequence embeddings, offering insights into both local and global properties of latent spaces. While methodological challenges, in particular hyperparameter sensitivity and comparison asymmetry, remain, addressing them could provide a way toward more biologically interpretable embeddings and deepen our understanding of what genomic models actually learn.

ATLASea : Challenges in building a comprehensive dataset in marine genomics

Poster

Isaline Guerin [1], ***Annie Lebreton*** [1], ***BYTE-Sea consortium*** [2], ***Erwan Corre*** [3]

1. *ABiMS bioinformatic platform, FR2424, CNRS/Sorbonne Université, Station Biologique de Roscoff (SBR),* ***2.*** *https://gitlab.com/pepr-atlasea/byte-sea_consortium,* ***3.*** *IFB-core, Institut Français de Bioinformatique (IFB), CNRS, INSERM, INRAE, CEA*

Abstract

The number of available marine genomes is rapidly expanding. However, marine biodiversity remains largely unexplored, with an estimated 90% of marine species still undescribed. Despite recent sequencing efforts, such as those conducted within the ATLASea programme (PEPR 2023–2031), only a limited number of high-quality reference genomes are publicly accessible, and an even smaller proportion have comprehensive and harmonised gene annotations. Furthermore, strong sampling biases and technical constraints — including access to organisms and the ability to obtain sufficient quantities of high-quality DNA — disproportionately favor particular taxonomic lineages. As a result, current genomic repositories provide an uneven and heterogeneous representation of marine diversity, further compounded by variability in annotation quality and database-dependent functional inference biases.

In this context, our objective is to construct a first overview of the structural and functional characteristics of public marine genomes and newly generated ATLASea genomes and to link this information to the existing environmental metadata, in particular that generated by the Tara Oceans expeditions.

Here we discuss the challenges involved in creating a comprehensive marine genomics dataset. We question the scope of high-quality marine genomes and the taxonomic biases encountered. We address challenges related to data accessibility, interoperability and computational costs associated with their exploitation.

ATLASea BYTE-Sea: Navigating IT Systems and Web Portals for Sample Tracking and Marine Data Exploitation

Poster

Loraine Brillet-Guéguen [1], ***Lucile Jeusset*** [2], ***Victor Leguet*** [3], ***Wael Ben Ammar*** [4], ***Alexandre Nicaise*** [5], ***Yaëlle Pihan*** [3], ***BYTE-Sea consortium*** [6], ***Erwan Corre*** [4]

1. *CNRS-Sorbonne University, Station Biologique de Roscoff, FR2424, ABiMS-IFB, Roscoff, France ; Sorbonne Université, CNRS, Integrative Biology of Marine Models (LBI2M), Station Biologique de Roscoff, Roscoff, France,* ***2.*** *Equipe DYOGEN, Institut de Biologie de l'ENS (IBENS), Département de biologie, École normale supérieure, CNRS (UMR8197), INSERM (U1024), Université PSL, 75005 Paris, France,* ***3.*** *Ifremer, IRSI, SeBiMER Service de Bioinformatique de l'Ifremer, F-29280 Plouzané, France,* ***4.*** *CNRS-Sorbonne University, Station Biologique de Roscoff, FR2424, ABiMS-IFB, Roscoff, France ; IFB-core, Institut Français de Bioinformatique (IFB), CNRS, INSERM, INRAE, CEA, 94800 Villejuif, France,* ***5.*** *CNRS-Sorbonne University, Station Biologique de Roscoff, FR2424, ABiMS-IFB, Roscoff, France,* ***6.*** *https://gitlab.com/pepr-atlasea/byte-sea_consortium*

Abstract

The ATLASea programme (PEPR 2023–2031, www.atlasea.fr), coordinated by CNRS and CEA, aims to sequence the genome of 4,500 marine species sampled along French coasts and overseas territories. To ensure the dissemination, interoperability, and accessibility of these genomic resources, the BYTE-Sea project, led by the French Bioinformatics Institute (IFB), has developed a robust digital infrastructure.

At its core, a centralized database integrates sampling metadata, sequencing status, and genome assembly metrics, serving as the backbone for three complementary web portals. The Tracker Portal, with restricted access, enables real-time monitoring of collected samples and sequencing progress, specifically designed for project stakeholders and decision-makers. The public Data Portal offers open access to comprehensive resources, including sequencing status, sampling metadata, genome assembly data, and species-specific genome browsers, thereby promoting transparency and enabling data reuse. Finally, the Marine Genomics Portal, currently under development, will centralize genomic resources from ATLASea and public databases, offering advanced tools for marine genome analysis, comparison, and visualization.

By integrating these platforms, BYTE-Sea creates a seamless workflow that supports efficient sample tracking, collaborative research, and data-driven discovery, fully aligned with FAIR principles and Open Science objectives.

URL

https://portal.atlasea.fr

Augmentating Pangenome Variation Graph With Low-coverage Sequencing for Haplotype Inference

Poster

Julien Chevreau [1], ***Camille Carrette*** [2], ***François Sabot*** [3], ***Christine Tranchant-Dubreuil*** [3]

1. *Université de Rouen,* ***2.*** *Université de Montpellier,* ***3.*** *IRD montpellier*

Abstract

Current linear reference genomes are the basis for many bioinformatics analysis, including variant calling or phylogenetic analysis. However, linear references introduce bias: sequences absent from the reference but present in new samples may not be studied. To overcome this limitation, Pangenome Variation Graphs (PVGs) enable to capture the full spectrum of genetic diversity within a species, from single nucleotide polymorphisms to large variations. PVGs have applications in crop improvement,ecology, and population genetics.

Although many tools have been developed to build and manipulate PVGs, graph editing and updating remain an active area of research. Currently, PVGs are typically built from genome assemblies, which requires expensive high-coverage sequencing to ensure a reliable graph structure. Graph augmentation could instead leverage the massive amount of available low-coverage sequencing data to improve graph completeness. Furthermore, an augmentation strategy could enable haplotype inference from new samples using cheap low-coverage sequencing. One possible starting point for such an approach is read-to-graph alignment [Figure 1]. To identify suitable tools for this step, we conducted a benchmark of existing tools.

We evaluated three aligners (vg Giraffe[1], GraphAligner[2], Minigraph[3]), a long-read augmentation tool (PALSS[4]) and a haplotype inference tool (PHI [5]) using both long and short reads on an African rice PVG. Considering specificity, sensibility, and performance, vg Giraffe proved to be the most suitable software for aligning both long and short reads and was chosen for subsequent analysis.

As PVG augmentation remains poorly explored, we propose three ways to augment a graph with different impacts on graph topology [Figure 2]. We focused on PVG augmentation to infer the haplotype of new samples. This scenario typically occurs in QTL studies involving Recombinant Inbred Lines for instance. Tests using simulated data are ongoing and future work will explore additional augmentation strategies, including the insertion of new nodes.

Automated construction of Boolean models using knowledge graphs

Poster

Nina Alger [1]***, Elisabeth Remy*** [2]***, Benno Schwikowski*** [3]***, Matthieu Najm*** [3]

1. *Université de Toulouse, Master Biologie Santé Parcours Complex Systems In Life Science (CSILS),* ***2.*** *Aix Marseille Univ, CNRS, I2M, Marseille, France,* ***3.*** *Institut Pasteur, Computational Systems Biomedicine*

Abstract

Boolean modeling has emerged as a powerful qualitative formalism to study the dynamics of gene regulatory networks, and in particular in cancer. A Boolean model (BM) consists of a prior knowledge network (PKN) — a signed directed graph capturing regulatory interactions — coupled with a set of logical rules, called network parametrization, that govern the activation state of each node. By simulating these dynamics, BMs predict the network's evolution through various functional states in response to different environmental conditions.

However, constructing the PKN remains a time-consuming and largely manual process due to the large amount of heterogeneous data available. Here, we present an automated pipeline for PKN construction and BM inference (see Figure).

First, a knowledge graph (KG) is assembled by integrating multiple prior knowledge databases using OntoWeaver and BioCypher: protein-protein interactions and pathways (OmniPath), drug-target interactions (OpenTargets), cancer gene biomarkers (OncoKB), and tissue-specific gene expression (Human Protein Atlas). Each database is integrated via dedicated Ontoweaver adapters, and BioCypher harmonizes the chosen ontology. Then, a context graph is extracted from the KG with user-defined biological queries — yielding a subnetwork relevant to the biological question. Tools such as NeKo are then used for network parametrization thereby building the PKN for the BM.

The approach is benchmarked against the manually constructed BM of Flobak et al. (2015) on AGS gastric cancer cells, assessing whether the automated network recapitulates the same drug synergy predictions. Automating PKN construction is a step toward personalized BM integrating patient-specific omics data.

URL

https://github.com/nina-alger/KG2BM

https://github.com/oncodash/ontoweaver

Automated structural annotation of marine eukaryotic genomes in the ATLASea project

Poster

Khaoula Ziane [1], ***Jean-Marc Aury*** [1], ***Benjamin Noel*** [1]

1. *Génomique Métabolique, Genoscope, Institut de Biologie François-Jacob, CEA - CNRS - Univ. Evry / Université Paris Saclay, 91000 Evry, France.*

Abstract

The ATLASea project aims to sequence, assemble, and annotate the genomes of thousands of marine eukaryotic species, covering a broad taxonomic diversity that includes mollusks, annelids, echinoderms, and fish. Producing reliable structural annotations at this scale requires a robust and largely automated workflow. Here, we describe a pipeline that integrates multiple sources of evidence, including protein databases, public and internal RNA-seq data, and ab initio gene predictors, within a unified gene model reconciliation framework. Applied so far to over a hundred marine eukaryotic genome assemblies, this workflow consistently generates high-quality annotations. Importantly, all annotations undergo systematic validation and quality assessment, including evaluation of gene completeness, structural consistency, and cross-evidence support, ensuring confidence in the resulting gene models and providing a solid foundation for large-scale comparative genomics of marine biodiversity.

Automatic characterization of regulatory elements in the human genome using multimodal integration of '-omics' data

Poster

Julien RAYNAL [1], ***Laurent BRÉHÉLIN*** [1], ***Charles LECELLIER*** [1]

1. *CNRS*

Abstract

Deciphering the DNA cis-regulatory code—critical for gene expression regulation—remains a major challenge in genetics and cancer research. While deep learning and statistical models now predict gene regulation and expression from DNA sequences, they primarily rely on dominant signals often located at promoters and enhancers. Yet, key regulatory features (e.g., open chromatin, transcription factor binding sites, and disease-associated variants) also reside outside these regions, where experimental data are sparse or noisy. As a result, current models struggle to assess the functional impact of genetic variants in these understudied areas. Building on our previous work demonstrating the predictability of transcription beyond canonical regions [Grapotte et al., Nat. Comm 2021], we aim to move beyond epigenetic segmentation (e.g., enhancers/promoters) by developing automated functional annotation methods. These methods leverage machine learning models optimized for specific genomic contexts. Here, I present models predicting CAGE-based transcription in unannotated regions and highlight their divergence from promoter- or enhancer-trained models. I then introduce a clustering approach to group regions by shared features (using different embeddings) and assign each cluster an optimal predictive model. This strategy promises genome-wide variant effect assessment, addressing the limitation of models restricted to canonical regulatory regions.

Automatic Mapping of UnLabelled Extracellular Transcripts (AMULET) for sparse spatial transcriptomics data

Poster

Gabriel Duval [1], ***Marcello Zago*** [1], ***Manfred Claassen*** [1]

1. *Universitätsklinikum Tübingen*

Abstract

Spatial transcriptomics technologies such as Xenium 5K and others are rapidly emerging as powerful tools for studying cellular interactions at the sub-micron resolution. However, one major limitation of these platforms is the extremely low number of transcripts captured per cell, often representing less than 1% of the expected transcriptome. This sparsity can and has been partially compensated by integrating data modalities with higher-coverage such as single-cell RNA-seq. This approach is not always feasible, particularly for poorly characterized tissues or highly heterogeneous diseases such as cancer.

Here, we propose to address this problem by increasing the read coverage of spatial transcriptomics data leveraging unassigned transcripts (those falling outside cell segmentation boundaries). In Xenium 5K data for instance, these can account for anywhere between 10% and nearly 60% of all detected transcripts. Although some are truly free-floating (e.g : extracellular RNA), many intracellular transcripts are misattributed due to segmentation errors or transcript diffusion artifacts. We propose recovering and reassigning these transcripts to their cells of origin and thereby improving transcriptomic coverage without requiring additional experimental data. To this end, we utilize an existing variational autoencoder (VAE)-based framework to denoise cellular gene expression and use this representation to systematically map unassigned transcripts back to their most probable cell of origin. Our approach leverages a combination of spatial features - including distance to neighboring cells, cell segmentation features, and local gene expression profiles - to probabilistically reassign transcripts in a biologically informed manner. Using simulated ground-truth datasets mimicking transcript diffusion, we correctly recover over 50% of extracellular transcripts. By recovering this signal, our method has the potential for more productive downstream analyses in spatial transcriptomics experiments.

URL

https://github.com/claassenlab/AMULET

Automating image-based severity assessment of watermelon mosaic virus symptoms in melon using deep learning

Poster

Matthieu Deloget [1], ***Jocelyn De Goer*** [2], ***Jacques Lagnel*** [3], ***Lucie Tamisier*** [3]

1. *INRAE,* ***2.*** *Université Clermont Auvergne, INRAE, VetAgro Sup, UMR EPIA, 63122 Saint-Genès-Champanelle,,* ***3.*** *INRAE, UR 1052 GAFL*

Abstract

Watermelon mosaic virus (WMV) is one of the most important viruses infecting cucurbits such as melon. In the absence of curative treatments, breeding genetically resistant varieties remains one of the most effective control strategies. This requires large-scale evaluation of symptom severity across diverse genotypes, a process currently performed manually by experts, making it time-consuming and observer-dependent. Here, we present an ongoing project aiming to develop an automated image-based system for scoring WMV symptom severity on melon, designed to provide standardized and reliable phenotypic data.

Our dataset comprises 1 542 curated melon leaf photographs acquired on a black background under controlled lighting and fixed camera angle, and 1,000 whole-plant photographs taken indoors with cluttered backgrounds and variable viewing angles. Leaf symptoms are classified into four ordinal severity classes (0: visually healthy; 3: severe symptoms), with a class distribution of 346/400/507/199. The proposed pipeline consists of two stages: a YOLO-based convolutional neural network for leaf detection and localization, followed by a classification CNN (ResNET50), to assess symptom severity on each detected leaf. Data augmentation — including 60° rotations, vertical reflections, and luminosity and contrast variations — is applied to improve generalization. Whole-plant severity is determined by the highest severity score among its detected leaves, consistent with standard phytopathological practice where the most symptomatic organ drives resistance assessment.

The system is designed to be deployed either via smartphone or through edge computing on a Raspberry Pi 5 equipped with a camera module and an AI HAT+, enabling in-situ phenotyping. Our target is to achieve at least 90% accuracy in severity classification. Preliminary results on leaf classification show promising classification performance and will be presented. Ultimately, this tool aims to support the identification of WMV-resistant melon genotypes and facilitate targeted breeding strategies.

Balancing Open Science and Data Privacy: The Challenge of Human Microbiome Research

Poster

Guillaume GAUTREAU [1], ***Aïcha EL JAI*** [2], ***Nicolas PONS*** [3], ***Cloud4SAMS Consortium*** [4], ***Claudine MEDIGUE*** [5], ***Hélène CHIAPELLO*** [2], ***Nathalie GANDON*** [6]

1. *Université Paris-Saclay, INRAE, MaIAGE,* ***2.*** *Université Paris-Saclay, INRAE, MaIAGE ; Université Paris-Saclay, INRAE, BioinfOmics, MIGALE bioinformatics facility, 78350 Jouy-en-Josas, France ; IFB-Core, IFB, CNRS, INSERM, INRAE, CEA,* ***3.*** *Université Paris-Saclay, INRAE, MGP,* ***4.*** *France 2030, PEPR SAMS,* ***5.*** *IFB-core, Institut Français de Bioinformatique (IFB), CNRS, INSERM, INRAE, CEA ; CNRS UMR8030, Université Evry-Val-d'Essonne, CEA, Genoscope, LABGeM ;,* ***6.*** *CODIR, Unité d'appui au Collège de direction, INRAE*

Abstract

In the era of data-driven research, the scientific community is navigating a complex tension between Open Science, which promotes data sharing to ensure transparency, reproducibility, and reusability, and Personal Data Protection, which safeguards individual privacy. Balancing these two principles is challenging in health sciences, where biological datasets are inherently linked to human identity in ways that are difficult to obscure. In human microbiome research, the open sharing of metagenomic read datasets has become common practice. Since 2010, >100k human metagenomes have been deposited in international open-access databases, and millions more are expected in the coming years. Meanwhile, growing scientific evidence has shown that microbiome composition cannot be regarded merely as an anonymous collection of microbes and may constitute a specific biological signature. Unlike the human genome, metagenomes may also reveal sensitive relational information, such as shared households or even social networks within communities. Because the microbiome is linked to health, publicly available sequences may be used to infer an individual's past, present, or future medical conditions, ranging from metabolic disorders to neurodegenerative diseases.

Consequently, there is a growing tendency to consider human metagenomes as data that may pose a risk of indirectly identifying personal health information. In the context of the Cloud4SAMS project, this poster aims to highlight the relevant scientific knowledge on this topic, discuss the resulting legal implications, and explore solutions to reconcile the dissemination of scientific results and data with the protection of personal privacy. We will also explain how the FAIR principles will be adapted to implement best practices throughout the human microbiome data lifecycle. This would ensure both security and privacy while preserving the possibility of reusability under defined conditions. Finally, we will present the first components of training materials that are currently being developed in the Cloud4SAMS project on this topic.

Bioinformatic development for Nanopore epigenomics: building reproducible workflows for methylation analysis and beyond

Poster

Laure FERRY [1], ***Mélina Farshchi*** [1], ***Magali Hennion*** [1]

1. *CNRS*

Abstract

Oxford Nanopore Technology (ONT) broadens epigenomic analysis by enabling direct detection of DNA modifications on native long reads, including 5-methylcytosine (5mC), 5-hydroxymethylcytosine (5hmC), and N6-methyladenine (m6A). These epigenetic marks regulate gene expression, help maintain cellular identity, and contribute to genome stability. Because long reads can span difficult genomic regions, ONT provides access to loci that are poorly resolved by short-read approaches. Combined with adaptive sampling, it also allows targeted enrichment without PCR amplification while preserving native modification signals.

Since Whole-Genome Bisulfite Sequencing (WGBS) remains the reference method for DNA methylation analysis, a major objective was to determine whether ONT could provide comparable methylation measurements. To support this, our team developed Methylator, a Snakemake/Apptainer pipeline designed for reproducible processing of both WGBS and ONT methylation data. Methylator integrates methylation extraction, CpG- and tile-level quantification, differential methylation analysis, and genomic annotation. It was also extended with functions for restricted-region filtering, coverage enrichment analysis, and harmonized processing across technologies, allowing ONT and WGBS datasets to be compared on identical genomic intervals.

Beyond this comparison framework, the Epigenetics and Cell Fate unit (UMR7216, Paris) hosts two complementary platforms for which this work was carried out: the BiBs platform (Bioinformatics and Biostatistics) and the EpiG platform (Epigenetics). Together, they offer wetlab and bioinformatic services for epigenome analysis to the research teams of the unit and beyond and participate in the development of new techniques. Supporting the diverse biological questions arising from these teams naturally requires dedicated bioinformatic developments. Rather than applying a single standard workflow, analysis strategies had to be adapted to the structure and objective of each experiment. This led to the implementation of pipelines for joint 5mC/5hmC analysis, region-specific variant calling and genotyping, and low-frequency variant detection in amplicon sequencing data, together with standalone analyses for questions not covered by existing tools.

URL

https://methylator.readthedohttps://methylator.readthedocs.io/en/latest/cs.io/en/latest/

https://github.com/parisepigenetics

BioloGrist: Using Grist for Biological Data Management - From Field Samples to Submission of Associated Sequencing Data

Poster

Laurent Brottier [1], ***Dalia Belmadi*** [1], ***Ania Saidani*** [1], ***Florence Auguy*** [1], ***Hamza Bouzayen*** [1], ***Stéphane De Mita*** [1], ***Sébastien Ravel*** [1], ***Sébastien Cunnac*** [1], ***Juliette Hayer*** [2], ***Alexis Dereeper*** [3]

1. *PHIM, CIRAD, INRAE, IRD, SupAgro, Université de Montpellier, F-34398 Montpellier, France,* ***2.*** *MIVEGEC, Université de Montpellier, IRD, CNRS,* ***3.*** *IRD, PHIM, South Green platform*

Abstract

At IRD in Montpellier, we developed a database based on the software Grist (https://www.getgrist.com/) that ensures traceability of data related to pathogen collections studied in our research units: from sampling data (GPS location, collection dates, sample photographs, regulatory information, …) to metadata associated with assembled genomes.

This work provides a flexible system that (i) facilitates data entry by research team members and their partners, (ii) enables the generation of synthetic dashboards summarizing the available data, and (iii) ensures data traceability and regulatory compliance while providing the foundation for data management plans (DMP).

Initially designed for the management of collections of phytopathogenic bacteria of the genus *Xanthomonas*, this solution is now used more broadly within the PHIM research unit to manage data related to other pathogens (notably fungi), both for collection management and, more generally, for the epidemiological monitoring of plant diseases.

As part of the « CLAPAS BioloGrist » project, implemented in 2025 through a collaboration with the MIVEGEC research unit, the objective was to consolidate this prototype and make it more generic so that it can also be applied to animal and human pathogens. The actions undertaken included connecting the tool KoboToolbox to the Grist database to enable field data entry directly from tablets, as well as developing a semi-automated protocol for submitting microbial genome sequencing data to public repositories such as European Nucleotide Archive (ENA) and Sequence Read Archive (SRA).

An IRD instance has been established to centralize the different Grist databases: https://biologrist.ird.fr/

URL

https://biologrist.ird.fr/

Bridging Scales: A Multi-Level Graph Neural Network for Protein Function Prediction

Poster

Antoine Toffano [1], ***Pierre Larmande*** [2], ***Jérôme Azé*** [1]

1. *LIRMM, Univ. Montpellier, CNRS,* ***2.*** *DIADE, Univ. Montpellier, IRD, CIRAD*

Abstract

The exponential growth of protein sequence data has outpaced experimental functional characterization, resulting in a widening annotation gap. Addressing this requires both the discovery of novel roles and the refinement of existing Gene Ontology annotations. Current computational approaches to this problem operate either at the atomic residue scale or the systemic network scale, which prevents a holistic understanding of protein functional roles. To address this limitation, we propose MS-GNN, a graph neural network framework that bridges these scales. At the protein level, we construct spatial graphs of amino acids through language model embeddings and AlphaFold-derived structural contact maps. At the systemic level, these representations are integrated into a global protein–protein association network. This unified architecture achieves state-of-the-art performance across all three sub-ontologies. Beyond architectural innovation, we demonstrate that complementing strict experimental labels with broader, curated functional data drastically improves performance, showing that annotation sparsity, rather than algorithmic capacity, is the primary bottleneck in model improvement.Finally, ablation studies confirm the necessity network-level information, particularly functional features, highlighting the necessity of multi-scale integration.

Bridging the gap in computational biology: genomic surveillance and bioinformatic innovation for plant health and food security in Africa

Poster

Justin S. PITA [1], ***Angela ENI*** [1], ***Fidele Tiendrebeogo*** [2], ***Ezechiel TIBIRI*** [3], ***Romaric K. NANEMA*** [4], ***Christine Tranchant-Dubreuil*** [5]

1. *Central and West African Virus Epidemiology (WAVE), Pôle scientifique et d'innovation de Bingerville, Université Félix Houphouët-Boigny (UFHB), Bingerville, Côte d'Ivoire,* ***2.*** *Central and West African Virus Epidemiology (WAVE), Pôle scientifique et d'innovation de Bingerville, Université Félix Houphouët-Boigny (UFHB),* ***3.*** *Laboratory of Virology and Plant Biotechnology, Institute of Environment and Agricultural Research (INERA), 01BP476,* ***4.*** *Genetic and Plant Breeding Team (EGAP), Biosciences Laboratory, Doctoral School of Science and Technology, Joseph KI-ZERBO University, Burkina Faso,* ***5.*** *IRD montpellier*

Abstract

In response to an increase in viral threats to staple crops in Central and West Africa, the WAVE (Central and West African Virus Epidemiology) network has developed a cross -border "One Health" approach that places bioinformatics at the heart of agricultural resilience. WAVE operates in 14 countries and uses High-Throughput Sequencing (HTS)—including Oxford Nanopore for in-field diagnostics and Illumina for deep characterization—to monitor viral evolution and emergence in real-time.

Central to our mission is the development and deployment of robust, reproducible bioinformatic pipelines tailored for viral metagenomics and phylogenomics. Utilizing workflow managers such as Snakemake, we ensure that complex genomic data analysis is standardized across the countries. These workflows facilitate the transition from raw signal to actionable epidemiological insights, utilizing High-Performance Computing (HPC) infrastructures to process large-scale datasets.

Beyond methodology, WAVE acts as a catalyst for bioinformatic capacity building by establishing an International Certificate in Bioinformatics and Genomics (CIBiG) (https://wave-centre.github.io/cibig/). By training a new generation of African scientists in computational biology, we are ensuring data sovereignty and the integration of data into Africa's development policies. This presentation will highlight how the integration of genomics, automated workflows, and international collaborations—such as the RABIAS (https://wave-centre.github.io/rabias/) network—is transforming plant health surveillance. By treating crop health as a fundamental pillar of the One Health triad, we demonstrate that bioinformatic sovereignty is the key to securing food systems and mitigating the socio-economic impacts of plant pandemics in the Global South.

URL

https://wave-centre.github.io/cibig/

Building a cassava pangenome to explore the genetic diversity of local cassava varieties from Côte d'Ivoire

Poster

Cyrielle Ndougonna [1], ***Christine Tranchant-Dubreuil*** [2], ***Ezechiel TIBIRI*** [3], ***Fidele Tiendrebeogo*** [4], ***Justin S. PITA*** [5]

1. *Central and West African Virus Epidemiology (WAVE),* ***2.*** *IRD Montpellier,* ***3.*** *Laboratory of Virology and Plant Biotechnology, Institute of Environment and Agricultural Research (INERA), 01BP476,* ***4.*** *Central and West African Virus Epidemiology (WAVE), Pôle scientifique et d'innovation de Bingerville, Université Félix Houphouët-Boigny (UFHB),* ***5.*** *Central and West African Virus Epidemiology (WAVE), Pôle scientifique et d'innovation de Bingerville, Université Félix Houphouët-Boigny (UFHB), Bingerville, Côte d'Ivoire*

Abstract

Cassava (*Manihot esculenta* Crantz) is the second-most important crop grown in Côte d'Ivoire after yam. While cassava is considered a climate-resilient crop[1], its production is at risk from whitefly-transmitted viruses such as cassava brown streak virus and cassava mosaic virus[2]. Breeding for resistance to pests and diseases and for tolerance to drought are therefore key priorities to sustain yields and ensure food security in the region. However, cassava breeding is hindered by the high degree of heterozygosity of the species and by the limited genomic resources available for this crop. Using a pangenome approach[3], our work aims to develop resources that can be used to explore the genetic diversity of cassava varieties grown in Côte d'Ivoire.

As a preliminary step towards constructing a pangenome variation graph (PVG) for African cassava varieties, we first prototyped a workflow to assess how the number of integrated genomes could impact the structure of the graph. Using 19 publicly available cassava genomes[4], we constructed a series of PVGs using Minigraph-Cactus[5] and assessed the impact of sequential genome addition and genome assembly quality on graph properties. Gene annotations from the reference genome AM560-2 (639.6 Mb) were transferred to the pangenome graph using GrAnnoT[6]. We then used GraTools[7] to extract key metrics, including graph size, node depth and core/dispensable genome ratio.

Our results highlight how genome assembly quality influences the resulting graph structure and quantify the increase in genomic diversity captured as additional genomes are incorporated. The metrics obtained were also compared with those reported in a recently published cassava pangenome study[8], providing a reference point to evaluate the consistency of our approach.

Building a Regional Bioinformatics Community in West Africa: Interdisciplinary Collaboration for Genomics and Health Research – RABIAS network

Poster

Julie ORJUELA [1], ***Ndomassi TANDO*** [1], ***Romaric K. NANEMA*** [2], ***Ezechiel TIBIRI*** [3], ***Christine Tranchant-Dubreuil*** [1], ***Justin S. PITA*** [4], ***Fidele Tiendrebeogo*** [5]

1. *IRD Montpellier,* ***2.*** *Genetic and Plant Breeding Team (EGAP), Biosciences Laboratory, Doctoral School of Science and Technology, Joseph KI-ZERBO University, Burkina Faso,* ***3.*** *Laboratory of Virology and Plant Biotechnology, Institute of Environment and Agricultural Research (INERA), 01BP476,* ***4.*** *Central and West African Virus Epidemiology (WAVE), Pôle scientifique et d'innovation de Bingerville, Université Félix Houphouët-Boigny (UFHB), Bingerville, Côte d'Ivoire,* ***5.*** *Central and West African Virus Epidemiology (WAVE), Pôle scientifique et d'innovation de Bingerville, Université Félix Houphouët-Boigny (UFHB)*

Abstract

Bioinformatics is a rapidly evolving interdisciplinary field that integrates biology, genomics, computer science, mathematics, and system administration. In West Africa, the need for a structured regional community has been increasingly recognized in recent years, particularly through training courses and capacity-building initiatives in Ivory Coast and Burkina Faso. To address this, the WAVE Center and IRD are organizing the First Interdisciplinary Meeting of the West African Bioinformatics Community (December 15 to 17, 2025, Abidjan), bringing together 40 scientists from 13 countries.

This meeting aims to establish a core network of bioinformaticians, system administrators, and researchers to foster interdisciplinary collaboration. Key objectives include promoting resource sharing, designing joint projects, and strengthening regional capacities in genomics and bioinformatics. The event features expert contributions from the French Bioinformatics Institute (IFB), the French Bioinformatics Society (SFBI), and Montpellier University, ensuring knowledge exchange and best practices. It brings together key stakeholders, including the African Society for Bioinformatics and Computational Biology (ASBCB), the Guinea Infectious Disease Research and Training Center (CERFIG), and the Afroscreen network, to discuss advancements in bioinformatics and genomic surveillance.

Initiatives such as H3ABioNet and WAVE (Central and West African Virus Epidemiology) are showcased, emphasizing capacity building, training, and research collaboration across the continent. The seminar also addresses the shared use of computational infrastructure and expertise, a priority for the West African System Administrators Network, established by the IRD and its African partners since 2015.

Expected outcomes include the creation of a collaborative framework for infrastructure sharing, capacity building, and funding opportunities. Long-term impacts involve improved recognition of bioinformatics as a discipline, enhanced regional collaboration, and stronger ties with international partners. This initiative represents a critical step toward advancing genomics and health research in West Africa.

URL

https://wave-centre.github.io///rabias//01.description.html

BYTE-Sea: Advances in the development of the digital infrastructure for ATLASea, the French marine genome sequencing programme

Poster

Annie Lebreton [1], ***BYTE-Sea consortium*** [2], ***Erwan Corre*** [3]

1. *ABiMS bioinformatic platform, FR2424, CNRS/Sorbonne Université, Station Biologique de Roscoff (SBR),* ***2.*** *https://gitlab.com/pepr-atlasea/byte-sea_consortium,* ***3.*** *IFB-core, Institut Français de Bioinformatique (IFB), CNRS, INSERM, INRAE, CEA*

Abstract

In the field of marine science, exploring diversity is a major challenge, as only 10% of marine species are known and few high-quality genomes of marine organisms have been made available to the scientific community. The ATLASea programme [1] (PEPR 2023-2031, led by the CNRS and the CEA) aims to decipher and exploit the vast wealth of information provided by marine biodiversity along the French coastline by sequencing 4,500 genomes of marine species collected in mainland France and French overseas territories. During the first three years of the programme, more than 7,000 samples corresponding to more than 2,000 distinct species across a wide range of phyla (26 in February 2026) were collected by the DIVE-Sea targeted project coordinated by the MNHN. More than 220 genomes have been assembled and published by the SEQ-Sea targeted project coordinated by the CEA. All genomic data produced by the ATLASea project are centralised on a single web portal [2] by the BYTE-Sea targeted project. BYTE-Sea, coordinated by the French Institute of Bioinformatics (IFB), guarantees data interoperability and security, and facilitates their dissemination and use in accordance with FAIR and Open Science principles. The various software environments developed provide access to marine organism genome sequencing data, including information on genomic sequence, genes, coding and non-coding regions, and will offer advanced tools for genome analysis, genome comparison and data visualisation.

URL

1. www.atlasea.fr
2. portal.atlasea.fr

Can somatic mutations be spatially localized using 10x Visium spatial transcriptomics?

Poster

sacha schutz [1]

1. *LBAI, UMR1227, Univ Brest, Inserm, France et (2) CHU de Brest, Brest, France*

Abstract

Somatic mutations are central drivers of tumor evolution and intratumoral heterogeneity. Identifying and mapping these variants is not only key to characterizing tumor biology, but also offers a unique opportunity to reconstruct the evolutionary history of cancer — tracing how distinct clones emerge, expand, and spatially organize within the tissue. While bulk and single-cell sequencing approaches have extensively catalogued mutational landscapes, they lack the spatial resolution required to position variants within their tissue context. Spatial transcriptomics platforms such as 10x Visium capture polyadenylated RNA from spatially barcoded spots, theoretically enabling the assignment of RNA-derived variants to precise tissue coordinates.

Methods. We analyzed a publicly available FF human pancreatic tissue dataset generated with the 10x Visium platform. Sequencing reads were aligned to the reference genome and assigned to their spatial barcodes, each carrying unique x,y tissue coordinates. Variant calling was performed on spatially resolved reads, followed by systematic filtering of germline variants to isolate high-confidence somatic candidates. Retained variants were annotated and projected onto the tissue section.

Results. At the minimum bin resolution of 8 µm, the number of reads per spatial unit was insufficient to reliably detect somatic variants, resulting in sparse and largely uninformative variant calls. However, as bin size was incrementally increased, read depth per bin grew accordingly, enriching the pool of detectable variants. By pooling reads across larger spatial bins, we recovered sufficient variant coverage to compute mutational signatures at a regional level. These signatures varied across the tissue section, suggesting the coexistence of heterogeneous mutational processes within spatially distinct regions.

Conclusion. Spatial transcriptomics data generated by 10x Visium can support somatic variant detection, despite coverage constraints. Through read-level spatial aggregation, mutational signatures can be reconstructed in situ, providing a new lens to study clonal dynamics and mutational processes within their microenvironmental context.

URL

https://github.com/dridk/space_variant

CARTOMIX: A generic web tool for the exploration of genome organization

Poster

Wolimata Diaw [1], ***Arthur Péré*** [1], ***Etienne G.J. Danchin*** [1], ***Marc Bailly-Bechet*** [1], ***Corinne Rancurel*** [2]

1. *Institut Sophia Agrobiotech (UMR1355), INRAE, Université Côte d'Azur, 400 route des chappes, 06903, Sophia Antipolis, France,* ***2.*** *PHYBAC (EMR7006), CNRS, INRAE, Université Côte d'Azur, 400 route des chappes, 06903, Sophia Antipolis, France*

Abstract

Genomic elements are not distributed randomly along chromosomes. They can have a biologically informative organization: genes with similar functions cluster together, genes, transposons, and non-coding RNAs can be co-localized in certain regions, and elements located in physical proximity can be subject to common regulatory mechanisms.

Exploring these organizations without prior hypotheses requires integrating and analyzing heterogeneous data from multiple sources and formats, including FASTA, GFF3, BED, and TSV. Existing tools only partially address this need: genome browsers offer excellent interactive visualization interfaces but limited statistical analysis, while command-line tools such as Bedtools are powerful but require programming skills. No tool currently combines visualization, statistics, and exploration for users without programming knowledge. This situation hinders scientists wishing to interrogate the spatial organization of their genomes.

Here we present Cartomix, an accessible, portable and reproducible web-based tool designed to facilitate the exploration of genomic spatial organization. Cartomix is structured around four main modules: multi-format data integration, genomic cartography and visualization, statistical computation, and user guidance to support exploratory choices. The application, deployed via Docker Compose, allows any user to upload their data, explore interactive visualizations (Circos plots) and export their results (figures and data).

Cartomix was tested on the genome of *Xiphinema index*, a phytoparasitic nematode and major vine pest with a 200 Mb genome divided into 10 chromosomes, for which a wealth of well-structured biological data is already available in our lab. This study will help address key questions for this organism, in particular regarding the distribution and co-localization of horizontally acquired genes and transposons. Beyond this use case, the Cartomix tool can be used on any sequenced genome and any combination of genomic data types.

cgMLST typing in the ABRomics web platform

Poster

Julie Lao [1], ***Raphaël Tackx*** [1], ***Amanda Dieuaide*** [1], ***Thomas Mignon*** [1], ***Cléa Siguret*** [1], ***Hugo Lefeuvre*** [2], ***Alix De Thoisy*** [3], ***Bérénice Batut*** [1], ***Nadia Goué*** [4], ***Sébastien Leclercq*** [5], ***Étienne Ruppé*** [6], ***Sylvain Brisse*** [3], ***Philippe Glaser*** [7], ***Claudine MEDIGUE*** [8], ***Fabien Mareuil*** [9]

1. *IFB-core, Institut Français de Bioinformatique (IFB), CNRS, INSERM, INRAE, CEA, 94800 Villejuif, France,* ***2.*** *Nantes Université, École Centrale Nantes, CNRS, LS2N, UMR 6004, F-44000 Nantes, France,* ***3.*** *Institut Pasteur, Université Paris Cité, Biodiversity and Epidemiology of Bacterial Pathogens, 75015, Paris, France,* ***4.*** *Plateforme AuBi, Mésocentre, Clermont-Auvergne, INRAE, UCA, 63000, Clermont-Ferrand, France,* ***5.*** *UMR ISP, INRAE, Université François Rabelais de Tours, 37380, Nouzilly, France,* ***6.*** *Université Paris Cité and Université Sorbonne Paris Nord, Inserm, IAME, 75018, Paris, France,* ***7.*** *Institut Pasteur, Université Paris Cité, Unité EERA, 75015, Paris, France,* ***8.*** *CNRS UMR8030, Université Evry-Val-d'Essonne, CEA, Genoscope, LABGeM, 91000, Evry, France,* ***9.*** *Institut Pasteur, Université Paris Cité, Bioinformatics and Biostatistics Hub 75015 Paris, France*

Abstract

ABRomics is a genome-based web platform for tracking multidrug-resistant bacteria and antibiotic resistance genes (ARGs). It automatically analyzes resistance content from submitted sequence data (FASTQ or FASTA files), and delivers online reports and Excel sheets containing metadata and analysis outputs (ARG/virulence detection, sequence type, plasmid typing, species identification, and quality control). Users can download individual files or retrieve multiple results for cross-sample analyses.

To track resistance transmission and foster collaboration, the platform provides a searchable database with filters for sample type, spatial/temporal metadata, antibiotic classes, resistance genes, species, and sequence types, alongside an interactive world map showing sample locations.

Since v1.3, ABRomics performs strain-level typing using core genome MultiLocus Sequence Typing (cgMLST) with CoreProfiler [1,2] and cgMLST schemes from BIGSdb-Pasteur [3], PubMLST [4], EnteroBase [5], and the cgMLST Nomenclature Server [6]. The platform enables simple cgMLST analysis using standardized nomenclatures from these four major resources. By linking public typing results with private strain collections, this approach offers an original One Health surveillance solution, allowing researchers to contextualize isolates across clinical, veterinary, and environmental sectors. Thus, integrating cgMLST, ARG detection, and geographic data enables high-resolution genomic epidemiology of multidrug-resistant pathogens from local to global scales.

1. https://gitlab.com/ifb-elixirfr/abromics/coreprofiler
2. Lefeuvre H, *et al.* github.com/iwc-workflows/cgmlst-bacterial-genome/main (v1.0). Zenodo. 2026. https://doi.org/10.5281/zenodo.18311510
3. Jolley KA, Maiden MC. BIGSdb: Scalable analysis of bacterial genome variation at the population level. *BMC Bioinformatics*. 2010;11(1):595.
4. Jolley KA, *et al.* Open-access bacterial population genomics: BIGSdb software, the PubMLST.org website and their applications. *Wellcome Open Res*. 2018;3:124.
5. Dyer NP, *et al.* EnteroBase in 2025: exploring the genomic epidemiology of bacterial pathogens. *Nucleic Acids Research*. 2025;53(D1):D757-D762.
6. https://www.cgmlst.org/ncs

URL

https://analysis.abromics.fr/

Characterization of grapevine fanleaf virus diversity and recombination events using complementary sequencing approaches.

Poster

Jeanne Juquel [1], ***Pierre Mustin*** [1], ***Jean-Michel Hily*** [2], ***Wassim Rhalloussi*** [1], ***Carine Schmitt*** [1], ***Myriam Hagege*** [1], ***Isabelle Rachel Martin*** [2], ***Olivier Lemaire*** [1], ***Anne Sicard*** [1], ***Emmanuelle Vigne*** [1], ***Sélim Ben Chéhida*** [1]

1. *INRAE, Université de Strasbourg, UMR-A 1131 Santé de la Vigne et Qualité du Vin,* ***2.*** *Institut Français de la Vigne et du Vin*

Abstract

Grapevine fanleaf virus (GFLV; *Nepovirus foliumflabelli*, *Secoviridae*) is a ssRNA(+) virus with a bipartite genome (RNA1 and RNA2), and the main agent of fanleaf degeneration disease in grapevines. Infected plants frequently harbor multiple GFLV variants resulting in mixed infections where recombination events could occur and contribute to viral genetic diversity and evolution. To investigate this diversity, a study was conducted on six infectious vineyard plots located in Burgundy and Champagne regions (France). Illumina sequencing was performed on total RNA extracted from a quarter of grapevines in each plot; 389 RNA1 and 351 RNA2 consensus sequences were thus obtained. Recombination analyses using Recombination Detection Program 5 software, identified several recombination hotspots across both genomic RNAs, allowing the identification of potential recombinant sequences and their putative parental variants. However, the short-read nature of Illumina sequencing (150 bp * 2) may complicate genome assembly in recombinant regions and potentially generate artefactual contigs after assembly. To validate the previously obtained consensus sequences, and thus the recombinant uncovered, complementary sequencing approaches have been employed. Samples exhibiting complex consensus sequences with evidence of recombination were selected for additional analyses using both Sanger and Oxford Nanopore Technologies (ONT) sequencing. Sanger sequencing was performed on amplicons generated with GFLV-specific primers flanking the identified recombination hotspots. In parallel, nanopore sequencing was evaluated using two library preparation strategies: (i) direct RNA sequencing and (ii) amplicon-based DNA sequencing following reverse transcription and PCR amplification targeting the ORFs of both genomic segments. Nanopore long reads were processed through a bioinformatics pipeline including basecalling and quality control with Dorado (ONT), *de novo* assembly with Flye , and polishing with Racon, to reconstruct complete GFLV genomes. The resulting assemblies have been compared with Sanger- and Illumina- derived sequences to assess the reliability of recombination detection and improve the characterization of GFLV genetic diversity.

Charting the Evolution of Protein Splice Variations Across the Tree of Life

Poster

Louis Carrel-Billiard [1]***, Arnaud Liehrmann*** [2]***, Hugues Richard*** [3]***, Élodie Laine*** [4]

1. *Department of Computational, Quantitative, and Synthetic Biology (CQSB), UMR 7238, IBPS,* ***2.*** *Department of Computational, Quantitative, and Synthetic Biology (CQSB), UMR 723,* ***3.*** *Genome Competence Center (MF1), Robert Koch Institute, Nordufer 20, 13353 Berlin, Germany,* ***4.*** *Department of Computational, Quantitative, and Synthetic Biology (CQSB), UMR 7238, IBPS, Sorbonne Université;Institut universitaire de France (IUF)*

Abstract

Alternative splicing fundamentally expands proteomic diversity, yet tracing the evolutionary history of individual splice events remains difficult: annotations are sparse outside model organisms, short exons are easily missed, and existing resources cover only a fraction of the tree of life.

We present SHIRE, a computational framework operating on evolutionary splicing graphs from ThorAxe combined with large-scale protein sequence resources such as omicsMSA and ColabFold. SHIRE maps homologous sequences across thousands of species and quantifies local conservation, enabling tracing of any type of alternative splicing event, whether a cassette exon inclusion/exclusion or mutually exclusive usage of tandem exons (MXE).

Applied to 121 human micro-exons and 235 tandem MXE events, events from both sets are recovered in evolutionary splicing graphs and extend across a broader taxonomic range than previously reported. Among the 121 micro-exons, 109 trace back to lineages far older than previously reported in the literature. The PTBP2 micro-exon, previously documented as conserved up to Bilateria, here extends to fungi, plants and protists, suggesting a eukaryotic origin over a billion years ago. Micro-exon conservation proves largely independent of host-protein divergence: some are highly conserved in taxa where the host protein has diverged substantially from the human ortholog.

Beyond validation of the tool on curated sets of alternative splicing events, we applied the framework to detect evolutionarily conserved alternative insertions and deletions across the human coding genome. This analysis highlighted NOVA1, a key splicing regulator in the brain, thereby illustrating how the limited taxonomic sampling of earlier studies can lead to overestimating the human specificity of molecular signatures.

Together, these findings reveal that splice-derived protein regions are older and more broadly distributed than current annotations capture, and enable systematic annotation transfer to uncharacterized lineages.

Cloud4SAMS: a trusted research environment to handle human gut microbiome data

Poster

Pauline BARBET [1], Eugeni Belda [2], Audrey Bihouée [3], Alexandrina Bodrug [3], Paul Breugnot [4], Stephane Delmotte [5], Guillaume GAUTREAU [6], Marie-Pierre Lasmenes [7], Rafael Patino-Navarrete [8], Brieuc Quemeneur [3], Matis Zouari [9], Cloud4SAMS Consortium [10], Frédéric Beck [4], Christophe Blanchet [11], Samuel Chaffron [12], Hélène CHIAPELLO [13], Karine Clément [8], Antoine Fabroulet [4], Alban Gaignard [3], Nathalie GANDON [7], David Salgado [14], Jacques van Helden [14], Claudine MEDIGUE [15], Nicolas PONS [1]

1. *Université Paris-Saclay, INRAE, MGP,* ***2.*** *IRD, Sorbonne Université, UMMISCO, F-93143,* ***3.*** *Nantes Université, CNRS, INSERM, l'Institut du Thorax,* ***4.*** *Inria, 78153 Le Chesnay,* ***5.*** *Universite Claude Bernard Lyon 1, LBBE, UMR 5558, CNRS, VetAgro Sup, Villeurbanne 69622, France,* ***6.*** *Université Paris-Saclay, INRAE, MaIAGE,* ***7.*** *CODIR, Unité d'appui au Collège de direction, INRAE,* ***8.*** *INSERM, Nutrition and Obesities; systemic approaches, NutriOmique, AP-HP, Hôpital Pitié-Salpêtrière,* ***9.*** *IFB-core, Institut Français de Bioinformatique (IFB), CNRS, INSERM, INRAE, CEA, 94800 Villejuif, France.,* ***10.*** *France 2030, PEPR SAMS,* ***11.*** *PRABI, Rhône-Alpes Bioinformatics Center, Université Lyon 1, 43 bd du 11 novembre 1918, Villeurbanne CEDEX 69622, France,* ***12.*** *Nantes Université, École Centrale Nantes, CNRS, LS2N, UMR 6004,* ***13.*** *Université Paris-Saclay, INRAE, MaIAGE, 78350, Jouy-en-Josas, France,* ***14.*** *IFB-core, Institut Français de Bioinformatique (IFB), CNRS, INSERM, INRAE, CEA,* ***15.*** *CNRS UMR8030, Université Evry-Val-d'Essonne, CEA, Genoscope, LABGeM, 91000, Evry, France*

Abstract

Analysing the influence of the microbiome on human health represents a scientific and medical challenge, raising both technological and ethical issues, such as the identification of data with predictive value for health. In this context, the objective of the Cloud4SAMS project is to deploy a Proof of Concept of a distributed digital infrastructure enabling scientists to exploit microbiome and health data in a secure computing environment. The Cloud4SAMS infrastructure will consist of resources (databases, software and analysis workflows) for processing datasets produced by scientific communities, in a secure computing and storage environment relying on the French Bioinformatics Institute (IFB) academic cloud federation (Biosphere) and the Health Data Storage (HDS) certified "secure bubble" (Arcana, Inria). Cloud4SAMS will bring support for the implementation of ethics and current regulations regarding the access, use and sharing of health and microbiome data. The architecture and implementation are based on the expertise and standards developed by the IFB (catalog of bioinformatics resources, FAIRification and data brokering tools) and by the ELIXIR network (bio.tools catalog, workflow hub, EDAM ontology, Federated European Genome Archive, Beacon…).

We describe here initial developments of the Cloud4SAMS architecture. Its entry point is a federated catalog of digital resources, which will list microbiome and health data, as well as software for their analysis. A module will manage requests to data access committees of each project, the validation of authorizations, and the ad hoc extraction and transfer of data. Deployment recipes will automate data orchestration with processing tools in a secure environment tailored to the data sensitivity level: either within HDS-type secure bubble for sensitive personal health-related data or within the IFB Biosphere cloud for non-sensitive data. Designed as a proof of concept, the feasibility and the performance of the infrastructure is evaluated through concrete use cases.

URL

https://www.ifb-elixir.fr/rd-innovation/projet/cloud4sams/

Community Detection in a Plant-based Fermentation Knowledge Graph

Poster

Zoé Le Roux [1], ***Alessandra Merlotti*** [1], ***Sandra Dérozier*** [2], ***Hélène CHIAPELLO*** [2], ***Daniel Remondini*** [1]

1. *University of Bologna,* ***2.*** *INRAE*

Abstract

Microbial data have rapidly expanded across diverse resources, including PubMed, omics datasets, and strain collections, creating a growing need for integration and standardization. To address this challenge, Omnicrobe was developed. It is a generalist database for microbes, their habitats, phenotypes, and uses. From this resource, we extracted data on plant-based fermentation. This study investigates whether the structure of this fermentation knowledge graph can be used to generate new hypotheses about fermentation processes. To this end, we performed community detection using the Louvain algorithm and the Barber modularity, which is adapted to bipartite networks. This article presents the first results on the features of the resulting communities.

Comparative analysis of regeneration transcriptomic landscape across animals

Poster

Yves CLEMENT [1], ***Eve Gazave*** [1]

1. *Institut Jacques Monod*

Abstract

Regeneration, the ability to restore body parts after amputation, is a phenomenon observed in most animal lineages. However, the regenerative potential varies drastically between species: in some, only specific organs can be regenerated (e.g. liver in humans or mice), while in others, the reformation of entire complex structures (e.g. limbs), or the whole body, can occur. Albeit these differences, all regeneration events are considered to follow three key steps: (i) wound healing, (ii) activation of precursors that will participate in the formation of a blastema (a regeneration specific structure, at the origin of the regenerated part), and (iii) morphogenesis.

Despite a long-lasting interest for regeneration, only a handful of comparative studies have tested a potential conservation of regeneration mechanisms in animals. However, these studies have important limitations: they focus on a handful of species, mostly vertebrates, and only a small number of regeneration types are sampled. As a consequence, whether there are conserved mechanisms or genetic programs governing regeneration in all animals remains an open and deeply important question. Our aim is to find if specific steps of regeneration are genetically conserved or divergent during evolution. Moreover, we aim at deciphering if the genetic machinery is conserved or divergent between major types of regeneration.

The number of RNA-seq studies focusing on regeneration in metazoan species has dramatically grown recently, allowing for detailed comparative studies in a variety of animals and the identification of conserved genes involved in regeneration across Metazoa. However, these comparative analyses are challenged by the complex history of genes across metazoans (duplications & losses) that make one-to-one comparisons scarce. We have adapted a comparative transcriptomic method to identify conserved gene expression clusters across large evolutionary distances that we successfully tested on embryonic data. We are currently applying this method on regeneration data.

Comparative Evaluation of Genomic Foundation Models for Regulatory Sequence Classification in Plant Genomes

Poster

Ibtissam Bouzidi [1], ***Mikael Lucas*** [1], ***Pierre Larmande*** [1]

1. *DIADE, Univ. Montpellier, IRD, CIRAD*

Abstract

Transcription factor binding sites (TFBS) are key regulatory elements controlling gene expression in plants, yet their genome-wide annotation remains limited in most non-model species. Here, we benchmark three genomic foundation models — AgroNT, BERT-TFBS, and Evo2 — for supervised TFBS classification across five plant species and a pooled multispecies dataset using frozen sequence embeddings. Evo2, trained on a broad genomic atlas spanning all domains of life without plant-specific emphasis, consistently achieves near-perfect classification (AUPRC $\geq$ 0.999) on four species, outperforming the plant-specialized AgroNT on shared species. Model scale does not consistently improve performance for short regulatory sequences: Evo2-7b outperforms Evo2-40b across all species. On the multispecies dataset, all models converge (AUPRC = 0.970–0.975). These results provide guidance on model selection for supervised TFBS classification and establish a pipeline extensible to genome-wide regulatory annotation.

Comparative genomics of phenotypic convergence and diversity in fishes

Poster

Alice REGNIER [1], ***Hugues Roest Crollius*** [2]

1. *Equipe DYOGEN, Institut de Biologie de l'ENS (IBENS), Département de biologie, École normale supérieure, CNRS (UMR8197), INSERM (U1024), Université PSL,* ***2.*** *Equipe DYOGEN, Institut de Biologie de l'ENS (IBENS), Département de biologie, École normale supérieure, CNRS (UMR8197), INSERM (U1024), Université PSL, 75005 Paris, France*

Abstract

In evolutionary biology, convergence refers to the independent evolution of similar morphological, physiological or behavioural traits in distinct lineages, reflecting adaptation to comparable ecological or functional constraints. Because convergent traits arise independently rather than from shared ancestry, they provide a powerful framework for disentangling phenotypic evolution from phylogenetic relatedness.

Across vertebrates, several studies have shown that phenotypic convergence is often associated with correlated genomic changes, including gene loss and shifts in evolutionary rates affecting coding sequences and regulatory elements. These genomic signatures reveal molecular mechanisms underlying the repeated emergence or loss of biological functions.

Fishes, which represent nearly half of all vertebrate species, display exceptional ecological and morphological diversity, with many traits evolving repeatedly across lineages. However, understanding how genomic composition and gene regulation contribute to these phenotypic patterns remains a major challenge.

To address this question, we combine comparative genomics, trait curation, and phylogenetically informed statistical analyses. We compiled a database of 314 high-quality fish genomes with gene annotations and curated trait information, including body size, habitat, electrogenicity, and air-breathing capacity. For each genome, we perform orthology inference and gene family size estimation using FastOMA and GLADE, to identify gene losses, gains, and duplication events in the context of the species phylogeny. These genomic changes will then be correlated with curated phenotypic data across fishes to detect gene families associated with the presence, absence, or modification of specific traits.

This framework will next allows us to investigate how genomic features such as gene content, evolutionary rate variation and regulatory changes are associated with the emergence, modification or loss of complex traits. By integrating phenotypic and genomic data across a broad phylogenetic scale, this project aims to uncover molecular signatures underlying morphological and ecological adaptations in fishes.

Comparing reference-based SNP analysis and k-mer approaches to assess genomic diversity of yam accessions from Burkina Faso within West African germplasm

Poster

SORY SIEDOU [1], ***DANSOU-KODJO Kodjovi Atassé*** [2], ***Christine Tranchant-Dubreuil*** [3], ***SCARCELLI Nora*** [4], ***Ezechiel TIBIRI*** [5], ***TIAMA Djakaridja*** [1], ***Fidele Tiendrebeogo*** [6], ***Romaric K. NANEMA*** [7]

1. *Centre National de la Recherche Scientifique et Technologique (CNRST), Institut de l'Environnement et de Recherches Agricoles (INERA), Burkina Faso.,* ***2.*** *International Certificate in Bioinformatics and Genomics (CIBiG),* ***3.*** *DIADE, University of Montpellier, CIRAD, IRD, 911 Avenue Agropolis, 34934 Montpellier Cedex 5, France,* ***4.*** *DIADE, équipe PANEEC IRD Montpellier 911 avenue d'Agropolis BP 64501 34394 Montpellier Cedex 5,* ***5.*** *Centre National de la Recherche Scientifique et Technologique (CNRST), Institut de l'Environnement et de Recherches Agricoles (INERA), Burkina Faso,* ***6.*** *Central and West African Virus Epidemiology (WAVE), Pôle scientifique et d'innovation de Bingerville, Université Félix Houphouët-Boigny (UFHB),* ***7.*** *Genetic and Plant Breeding Team (EGAP), Biosciences Laboratory, Doctoral School of Science and Technology, Joseph KI-ZERBO University, Burkina Faso*

Abstract

Yam (*Dioscorea rotundata*) is a major staple crop in West Africa and represents an important source of carbohydrates for rural populations. Despite its agronomic and socio-economic importance, yam genetic resources from Burkina Faso remain poorly characterized at the genomic level. In this study, we analyzed the genomic diversity of six local accessions of *Dioscorea rotundata* collected in Burkina Faso and integrated them into a reference dataset of 167 West African accessions from the study of Nora Scarcelli et *al.* (2019). High-throughput sequencing data were processed using a bioinformatics pipeline inspired by the original methodology but incorporating updated software tools for read alignment, variant calling, and stringent filtering of single nucleotide polymorphisms (SNPs). After filtering, approximately 3.57 million high-quality SNPs were retained for downstream population genomic analyses, including principal component analysis (PCA) and clustering methods. In addition to this reference-based SNP analysis, a reference-free approach based on k-mer composition was implemented to assess genetic relationships independently of the reference genome. The main objective of this study was to compare the results obtained from these complementary approaches in order to evaluate their consistency and their ability to describe yam genomic diversity. Population structure analyses revealed that one of the six Burkina Faso accessions separates from the others, while the remaining five cluster with cultivated yam accessions from West Africa. Overall, the two analytical strategies produced consistent patterns of genetic relationships. This study provides a first genomic insight into yam genetic resources from Burkina Faso and highlights the value of combining reference-based and reference-free approaches for population genomic analyses.

Computational deciphering and mathematical modeling of the regulatory networks controlling plasmacytoïd dendritic cell biology

Poster

Arafate IDRISSOU [1], ***Lucie Lamothe*** [2], ***Laurent HANNOUCHE*** [1], ***Bertrand ESCALIERE*** [1], ***Clemence GARREC*** [1], ***Marine ZAFFRAN*** [1], ***Laurine GIL*** [1], ***Lea David*** [1], ***Camille PIERINI-MALOSSE*** [1], ***Jean DESCAMPS*** [1], ***Pierre MILPIED*** [1], ***Elena TOMASELLO*** [1], ***Lionel SPINELLI*** [1], ***Magali Richard*** [2], ***Marc DALOD*** [1]

1. *Aix-Marseille Univ, CNRS, INSERM, CIML, Turing Centre for Living Systems, 13009 Marseille, France,* ***2.*** *Univ. Grenoble Alpes, CNRS, Grenoble INP, LIG, 38000 Grenoble, France*

Abstract

Plasmacytoid Dendritic Cells (pDCs) are specialized immune cells that are proposed to play a key role in antiviral defense and also promote several autoimmune diseases, through their ability to produce high levels of type I and III interferons (IFN-I/III)[1]. However, their identity and other functions remain debated, due to their remarkable phenotypic and functional plasticity. Phenotypically, pDCs express distinct gene modules, some of which are specific to their lineage, while others are shared with other Dendritic Cell (DC) subsets or with Innate Lymphoid Cells (ILCs) [2,3]. Our recent study has shown that, in vivo in mice infected with cytomegalovirus, pDCs undergo distinct activation states that are associated to specific functional potentials, subsequently in time and in different micro-anatomical locations [4]. Our project aims first to better understand the identity of pDCs, by inferring the nature of the relationships of pDCs with other DCs and ILCs, based on the proximity of their gene expression profiles, the activity status of gene regulatory networks, and the associated functional annotations. To this aim, we are building and analyzing an atlas of different types of immune cells isolated from the murine spleen. The resulting hypotheses will then be tested experimentally. We also want to determine how the expression of the gene modules of pDCs varies between their successive activation states and according to their tissue of microanatomical location.

Computational prediction of transcription factor binding to DNA using deep learning

Poster

Agathe Bancquart [1], ***Anaïs Bardet*** [1]

1. *IGBMC*

Abstract

Cellular heterogeneity largely results from differences in transcriptional states that control gene expression. The establishment and maintenance of these states rely on the cooperativity within a family of regulatory proteins known as transcription factors (TFs). These proteins recognize and bind to short DNA motifs located at variable distances from their target genes. Since their motifs are very short, millions of occurrences can be found throughout the genome. However, only a few thousand of these sites are actually bound, and their occupancy changes dynamically across cell types and conditions.

To identify transcription factor binding sites, experiments such as ChIP-seq are commonly used, but they present limitations that restrict their application. Broader studies of TFs binding sites in a given cell type can be carried out by measuring chromatin accessibility (ATAC-seq) but do not distinguish between individual TFs. To address these limitations, bioinformatics approaches recently based on deep learning models, have been developed, using either DNA sequence alone or in combination with chromatin accessibility data that include cell types specificities. Thus, the goal of my project is to develop a deep learning approach to precisely predict TF binding sites in a cell type specific manner for as many TFs as possible.

We developed a convolution neural network (CNN), which takes as input a DNA sequence combined with the chromatin accessibility signal, and predicts probabilities to have a TF binding site each 32bp resolution in the given cell type.

Our initial results show that more advanced data curation using stringent criteria significantly improves the quality of our predictions. These promising outcomes highlight the importance of data quality and provide a solid foundation for outperforming current models by refining both the architecture and the training strategy.

Computationnal approaches for studying readthrough transcripts biogenesis and functions in neuroblastoma cells

Poster

Lou-Sahra Khourab [1], ***Khouaila Aouadi*** [1], ***William DESAINTJEAN*** [1], ***Alizée DUQUET*** [1], ***Hélène Polvèche*** [1], ***Franck Mortreux*** [1], ***Cyril Bourgeois*** [1]

1. *Laboratoire de Biologie et Modélisation de la Cellule, École Normale Supérieure de Lyon, CNRS, UMR 5239, Inserm, U1293, Université Claude Bernard Lyon 1*

Abstract

Correct transcription termination, which relies on the recognition of the transcription termination site (TTS), is an essential step in gene expression. Under cellular stress or in diseases, the TTS is occasionally not recognized, allowing transcription to continue beyond its normal boundary (readthrough transcription). When two genes are genomically positionned in tandem, readthrough transcription can invade the downstream gene, generating transcription readthrough-associated chimeric RNAs (tracRNAs). The DEAD box helicases DDX5 and DDX17 are key regulators of RNA metabolism, transcription termination (Terrone et al., 2022, Nucleic Acids Res.) and RNA/DNA structure modulation. In neuroblastoma, where the MYCN oncogene is frequently amplified, the team showed that DDX17 and MYCN interact and that DDX5 and DDX17 depletion increases tracRNA production (Clerc et al., 2024, BiorXiv).

This project aims to elucidate the mechanisms linking helicases activity (especially on G-quadruplex) and readthrough transcription and to characterize the role of tracRNAs in neuroblastoma. Our mechanistic approach will integrate mutli-omics data, focusing on long read sequencing and G4 mapping to understand TTS recognition regulation. The functional part of the project will involve studying the non-coding functions of tracRNAs (as miRNA sponges or interacting with RNA binding proteins) and characterizing coding tracRNAs to assess their potential neoantigen production. Finally, we will develop a comprehensive computational atlas of tracRNAs which will be built from cell lines data and enriched with patient tumor data. The goal is to better characterize neuroblastoma-specific tracRNAs and provide a resource for identifying tracRNAs common to other cancer types.

URL

https://pmc.ncbi.nlm.nih.gov/articles/PMC9458439/
https://www.researchsquare.com/article/rs-7982361/v1

Could the methylome be a new lever for steering microbial communities?

Poster

Benjamin Prehaud [1], ***Iacopo Passeri*** [2], ***Joël Doré*** [3], ***Béatrice de Montera*** [4], ***Guillaume GAUTREAU*** [1]

1. *Université Paris-Saclay, INRAE, MaIAGE, 78350, Jouy-en-Josas, France,* ***2.*** *Department of Biology, University of Florence,* ***3.*** *University Paris-Saclay, INRAE, AgroParisTech, Micalis Institute; University Paris-Saclay, INRAE, Metagenopolis,* ***4.*** *UR CONFLUENCE: Sciences et Humanites (EA 1598), UCLy*

Abstract

DNA methylation is a critical epigenetic regulator in prokaryotes, influencing essential processes such as DNA replication, gene expression, and cellular defense. While metagenomics has traditionally focused on taxonomic composition (who is there) and functional potential (what they can do), the epigenetic layer remains a largely untapped source of information. Yet, long-read metagenomics now enables the characterization of the metagenomic methylome.

We present an automated pipeline for the integrated co-analysis of the microbial methylome, taxonomic abundance, and in situ growth rates. Using long-read sequencing data (PacBio or Oxford Nanopore), the workflow simultaneously reconstructs metagenome-assembled genomes (MAGs) and identifies base modification motifs (5mC, 5hmC, 4mC, and 6mA). The core innovation of this tool lies in its ability to correlate these epigenetic signatures with quantitative ecological metrics within a single cohesive environment.

Specifically, the pipeline integrates GRiD (Growth Rate Index) to estimate replication rates, determined by the coverage gradient between the origin and terminus of replication, alongside taxon-specific abundance. This multidimensional approach allows to identify specific methylation motifs that are statistically associated with high-growth phases or shifts in population structure. By mapping the epigenetic state of a taxon against its replication rate and relative abundance, the pipeline provides a new lens through which to investigate how methylation-mediated regulation influences the fitness and competitive advantages of taxa within an ecosystem.

Analyses of diverse metagenomic datasets will be presented, highlighting candidate epigenetic drivers correlated with microbial population dynamics. This framework is expected to offer a bioinformatic solution for exploring the epigenetic response of uncultured bacteria to environmental changes within a symbiotic context, providing new insights into the pressures shaping complex microbiotas. Ultimately, deciphering these epigenetic signatures, along with their transferases, could provide new levers to modulate population dynamics.

CurateMake: a reproducible multi-source workflow for ITS reference database curation in metabarcoding

Poster

Auguste Gardette [1], ***Eugeni Belda*** [1], ***Edi Prifti*** [1], ***Jean-Daniel Zucker*** [1]

1. *IRD, Sorbonne Université, UMMISCO, F-93143*

Abstract

Accurate taxonomic assignments in DNA metabarcoding depend on the quality of reference databases. For the ITS (Internal Transcribed Spacer) marker, the standard barcode for fungi and plants, major repositories (UNITE, BOLD, PLANiTS, CALeDNA) collectively harbour over 3.8 million sequences, yet remain fragmented, nomenclaturally inconsistent, and contain 5–20% in public databases. These errors propagate into downstream biodiversity analyses and may compromise ecological conclusions.

We present CurateMake, a reproducible Snakemake workflow that fuses, harmonises, and phylogenetically validates multi-source ITS reference databases. It combines nomenclatural harmonisation via the Catalogue of Life (CoL) API, ITS sub-region extraction (ITSx), hierarchical multiple sequence alignment (MAFFT/HMMER) with clade-partitioned tasks, phylogeny-aware label validation (SATIVA), and automated audit dashboards. Three parallel taxonomic backbones (Raw, CoL, Sativa) are preserved throughout for traceability.

We evaluated CurateMake with two complementary strategies. Intra-cluster Shannon entropy analysis (MMseqs2, 95–99% identity) shows a significant and consistent reduction across all taxonomic ranks (Wilcoxon signed-rank test), with the Sativa backbone presenting the lowest entropy, followed by CoL and Raw. Kingdom-level entropy reaches near zero in the Sativa backbone, indicating effective removal of cross-kingdom contaminations. A controlled error simulation using in silico full-length ITS sequence variants generated from 20 reference taxa (Carex spp., Russula spp.) with artificial annotation errors shows that CurateMake corrects 28% of errors at 15% corruption, outperforming CoL alone (~21%). SATIVA applied directly on a global alignment fails to produce meaningful corrections, as the hypervariable nature of the ITS region prevents the construction of reliable alignments at database scale. CurateMake overcomes this limitation through its hierarchical, clade-partitioned MSA strategy, and is the only approach capable of flagging contaminations.

CurateMake delivers an auditable ITS reference database and a methodological framework for quantifying annotation reliability in the absence of ground truth.

D-Genies2 : dot plot large genomes in an interactive, more efficient and simpler way.

Poster

Vincent Dominguez [1], ***Philippe Bordron*** [1], ***Christophe Klopp*** [1]

1. *Université de Toulouse, INRAE, UR 875 MIAT, F-31320*

Abstract

Third generation sequencing technologies enable to produce new, ever larger genomes assemblies. D-Genies is a tool used in this context. It allows, through a dotplot representation, to assess novel assembly quality by comparing it with existing references. It shows large modifications between assemblies such as insertions, deletions, inversions and translocations.

The current D-Genies uses flask library to generate and serve a web interface on the fly. The interface is a HTML form allowing to submit files in a job creating a dotplot. It uses D3.js for display and interaction with the doplots. D-Genies allows to run one job easily, but lacks ergonomy to compose job series.

D-Genies2 interface was designed to improve User experience (UX). The interface is based on the Vue.js framework and uses blocks manipulation to compose jobs. For example, dragging and dropping files (or url) on the interface will add those files to the file list. Then this list can be dropped on a job profile to generate a batch of jobs. Many job types are available to ease job composition. Jobs can be modified individually or in groups.

The display and interaction with the dotplot are also smoother, in particular for huge assemblies, thanks to the use of WebGL rendering instead of SVG rendering, and some calculations performed on the interface side rather than on the server side.

D-Genies2 adopts a client-server approach. The server no longer generates or serves pages. Instead, it provides an API to interact with. The web interface acts as a client, but it is also possible to submit jobs to D-Genies without using the web interface. A command line client is provided to interact with D-Genies from another computer like a HPC cluster.

D-Genies2 is an ongoing work, new features are expected to be added in the near future.

Data mining of public genomic repositories: harnessing off-target reads to expand microbial pathogen genomic resources

Poster

damien richard [1], ***Nils Poulicard*** [1]

1. *PHIM, CIRAD, INRAE, IRD, SupAgro, Université de Montpellier, F-34398 Montpellier, France*

Abstract

As sequencing technologies become more affordable and genomic databases expand continuously, the reuse of publicly available sequencing data emerges as a powerful strategy for studying microbial pathogens. Indeed, raw sequencing reads generated for the study of a given organism often contain reads originating from the associated microbiota. This review explores how such off-target reads can be detected and used for the study of microbial pathogens. We present genomic data mining as a method to identify relevant sequencing runs from petabase-scale databases, highlighting recent methodological advances that allow efficient database querying. We then briefly outline methods designed to retrieve relevant data and associated metadata, and provide an overview of common downstream analysis pipelines. We discuss how such approaches have (i) expanded the known genetic diversity of microbial pathogens, (ii) enriched our understanding of their spatiotemporal distribution, and (iii) highlighted previously unrecognized ecological interactions involving microbial pathogens. However, these analyses often rely on the completeness and accuracy of accompanying metadata, which remain highly variable. We detail common pitfalls, including data contamination and metadata misannotations, and suggest strategies for result interpretation. Ultimately, while data mining cannot replace dedicated studies, it constitutes an essential and complementary tool for microbial pathogen research. Broader utility will depend on improved data standardization and systematic genomic monitoring across ecosystems.

URL

https://peercommunityjournal.org/articles/10.24072/pcjournal.637/

Deciphering the photoperiod-driven life cycle of the non-model algae Tisochrysis lutea through Single-Cell Transcriptomics

Poster

Antoine Daussin [1], ***Laura PAGEAULT*** [2], ***Cyril NOEL*** [1], ***Laura LEROI*** [1], ***Gregory CARRIER*** [2], ***Bruno SAINT-JEAN*** [2]

1. *Ifremer, IRSI, SeBiMER Service de Bioinformatique de l'Ifremer, F-29280 Plouzané, France,* ***2.*** *Ifremer, PHYTOX, GENALG, F-44311 Nantes, France*

Abstract

For decades, clonal cultures of unicellular microalgae were considered genetically and functionally homogeneous. However, recent studies have revealed transcriptional and functional heterogeneity within these clonal populations [1]. This variability arises from multiple sources, such as spontaneous mutations in microorganisms, which can lead to sub-clones with distinct phenotypes. Nevertheless, most intercellular differences are not solely due to stable genetic modifications, but rather to stochastic fluctuations in biological processes, including gene expression noise. Furthermore, phenotypic variability also arises from environmental factors including nutrient availability and lighting conditions, shaping cellular states. Population-scale transcriptomic approaches often mask this diversity, making single-cell RNA sequencing (scRNA-seq) a powerful approach to capture gene expression profiles at the individual cell level [2,3].

In this study, scRNA-seq was applied to the microalga *Tisochrysis lutea* to investigate cellular heterogeneity and identify genes and/or subpopulations associated with photoperiod. Cultures were grown under light or dark conditions with two biological replicates. Cells were processed using the 10x Chromium and sequenced on Illumina, yielding 883M reads for ~20,000 collected cells.

Raw sequencing data were processed through a reproducible bioinformatic pipeline. Reads were aligned to the *T. lutea* v3 *de novo* genome assembly [4] using STARsolo [5]. The resulting gene expression matrices were analysed in R with Seurat [6] package, including quality control, low-quality cell filtering, normalization, dimension reduction to identify transcriptionally distinct cell populations through clustering and differential gene expression analysis. Because many algae remain poorly characterised, cell type annotation relied on protein domain prediction using InterProScan rather than orthology-based methods. Conserved domains were used to assign Gene Ontology categories and perform enrichment analyses of the identified clusters.

Resolving photoperiod-associated cell states would provide a framework to identify additional cell types or states in *T. lutea*.

The dataset, associated metadata, and analysis workflow follow FAIR data principles to allow reproducibility.

Deciphering translational regulation during infection with the Sindbis virus

Poster

Lauryn Trouillot [1], ***David Cluet*** [1], ***Emiliano Ricci*** [1], ***Christelle Morris*** [1]

1. *Laboratoire de Biologie et Modélisation de la Cellule, École Normale Supérieure de Lyon, CNRS, UMR 5239, Inserm, U1293, Université Claude Bernard Lyon 1*

Abstract

Viruses hijack host ribosomes to synthesize their proteins and have evolved various strategies to manipulate translation, including the association of numerous non-ribosomal proteins with the core ribosome. In a previous work, we identified a set of ribosome-associated proteins that display differential binding activity in response to infection with the alphavirus Sindbis. Several of these proteins are critical for viral propagation, including the helicase ASCC3. Our results suggest that ASCC3 acts independently of its canonical function of rescuing ribosome collisions. Instead, it promotes the synthesis of proteins containing a signal peptide.

We now aim to elucidate the role of ASCC3 in this newly identified pathway. RNA-seq and Ribo-seq data were generated from control cells and ASCC3-knockout cells that were uninfected or infected with Sindbis. Using a gradient-boosting machine learning approach, we identified several key protein and RNA features that can explain the positive or negative impact of ASCC3 down-regulation. Most of these features correlate with the presence or composition of a signal peptide. To further investigate these findings, we developed a ribosome profiling pipeline to validate the leads identified by the model. This bioinformatic pipeline first preprocesses ribosome profiling data, including size-based filtering steps to distinguish ribosome populations such as monosomes, disomes or multiple collided ribosomes. Reads are then aligned to the reference genome using STAR to quantify ribosome density along each transcript. Since viral infection is known to perturb translation initiation, we further investigate the potential emergence of upstream open reading frames (uORFs) and ribosome frameshifting events using RiboCode. Finally, translational efficiency is evaluated by computing the ratio of ribosome footprints to mRNA abundance for each transcript. The resulting features extracted from this pipeline will be reintegrated into the gradient-boosting model to refine our understanding of ASCC3's role in translation.

Deciphering virus-host-environment relationships guided by large scale metagenomics data integration: the Dziani Dzaha hypersaline lake virome case study

Poster

Maël Rimeur [1], ***Esther Mangelinck*** [2], ***Valentine Banneville*** [2], ***Aurore Wafflart*** [2], ***Mariama Drame*** [2], ***Christine Oger*** [3], ***Elea Pauliat*** [3], ***Paul Tissot*** [3], ***Mélodie Fleury*** [3], ***Laurence Josset*** [4], ***Jocelyn Turpin*** [5], ***Oldrich Navratil*** [6], ***Vincent Navratil*** [3], ***Mylène Hugoni*** [7]

1. *PRABI, Rhône-Alpes Bioinformatics Center, Université Lyon 1, 43 bd du 11 novembre 1918, Villeurbanne CEDEX 69622, France.,* ***2.*** *Students of the Biodiversity, Ecology, Evolution Master of Lyon, Villeurbanne 69100, France,* ***3.*** *PRABI, Rhône-Alpes Bioinformatics Center, Université Lyon 1, 43 bd du 11 novembre 1918, Villeurbanne CEDEX 69622, France,* ***4.*** *Hospices civils de Lyon,* ***5.*** *IVPC UMR754, INRAE, Universite Claude Bernard Lyon 1, EPHE, PSL Research University, 69007, Lyon, France,* ***6.*** *CNRS 5600 EVS, Université Lumière Lyon 2,* ***7.*** *Univ Lyon, INSA Lyon, CNRS, UMR5240 Microbiologie Adaptation et Pathogénie, F-69621 Villeurbanne, France. Institut Universitaire de France (IUF)*

Abstract

Background: Viruses are increasingly recognized as critical regulators of environmental dynamics, influencing microbial composition across ecosystems [1,2]. Most viruses remain unknown because they are difficult to cultivate or isolate. However, recent advances in viral metagenomics enable massive recovery of viral Metagenome Assembled Genomes (vMAG) from numerous ecosystems now distributed in new open archives [3,4]. The virosphere of extreme environments, such as hypersaline lakes remains poorly described. In this work, we focused on the hypersaline Lake Dziani Dzaha (Mayotte) virome. Previous studies characterized the bacterial, archaeal, and microeukaryotic communities [5,6], so the viral community and its interactions with the microbiome remain unknown. The aim is to determine virome spatio-temporal dynamics and relationships with the microbial community.

Results: We combined a state-of-the-art metagenomic pipeline with a planetary-scale approach to recontextualize viral contigs from extreme environments. We identified 378 viral high-quality contigs from the co-assembly of long term monitoring metagenomics. We highlighted a temporal dynamic in the composition of the lake virome, as well global scale similarity with other saline and hypersaline aquatic environments. Using reconstructed MAGs, several virus–host interactions were identified and are under investigation. Despite the growing number of viromic pipelines [7,8], few address virus–host interactions, and none integrate virus–host interactions with ecosystem dynamics. To address this limitation, we developed a user-friendly Snakemake-based pipeline that performs standard virome analyses and metadata visualization, helping mitigate limited taxonomic resolution for viral organisms.

Conclusions: This work illustrates the need to compare viral diversity with data obtained from other ecosystems at global scale. Our workflow aims to facilitate the exploration of virus-host interactions and viral dynamics in various ecosystems or hosts. Round trip between large-scale initiatives such as Virome@tlas and more local experimental setup will further improve the contextualization of viral diversity and understanding of its ecological roles.

DeconvolisSTa - Deconvolution of Spatial Transcriptomics dAta

Poster

Abderahim Lagraoui [1], ***Nejma Moualhi*** [1], ***Enola Missonnier*** [2], ***Maialen Arrieta*** [1], ***Slim Karkar*** [1]

1. *Université de Bordeaux,* ***2.*** *Universidad de Alicante*

Abstract

DeconvoliSTa is an open-source software tool designed for the deconvolution of sequence-based, RNA-Seq, spatial transcriptomics (ST) data. It estimates the proportion of cell-type-specific expression profiles of each spot as mini-bulk. Build on **Spotless** [1], the use of **Snakemake workflow manager (**instead of **NextFlow)** provides greater flexibility for adding methods, pre- and post-processing steps, and adapting pipelines to user's needs.

Based on the very comprehensive benchmark Spotless, the tool naturally integrates various deconvolution methods to estimate the cellular composition of heterogeneous tissue samples, utilizing single-cell reference datasets to improve accuracy. All deconvolution methods are readily available through their dedicated Docker image, ensuring a consistent computational environment. It also provides gold standards using simulated datasets and silver standards using image-based omics and segmentation to define ground truth, making it a convenient tool to generate and apply pipelines on user-defined simulated datasets.

Taking fully advantage of SnakeMake, users and developers can easily add new steps or implement additional deconvolution methods in R, Python or any Docker-based method and adapt to diverse experimental setups. For example, DeconvoliSTa offers now advanced visualization outputs. It allows for the comparison of results from various deconvolution tools or even ground truth on the same data. It features user-friendly navigation across the sample image and the superposed composition of the spot, enhancing clarity in analyzing cellular dynamics.

By giving access to several reference expression profiles and deconvolution methods, Deconvolista accurately infers and compares the contributions of individual cell types in Spatial Transcriptomics data in various domain as cancer biology, immunology, and developmental biology.

URL

https://github.com/cbib/DeconvoliSTa/

DeCovarT: Network-Driven Deconvolution of Transcriptomics data to dissect organoid Cellular Heterogeneity

Poster

Bastien Chassagnol [1], ***Anaïs Baudot*** [2], ***Grégory Nuel*** [2], ***Etienne Becht*** [3]

1. *Aix-Marseille Université, Marseille Medical Genetics,* ***2.*** *CNRS,* ***3.*** *INSERM*

Abstract

Bulk transcriptomics, while widely used to characterise biological systems, suffers from a fundamental limitation: gene expression measurements aggregate signals from heterogeneous cell populations, confounding cell-type-specific expression with variation in cellular composition. Single-cell technologies can resolve this but remain costly and technically demanding. Deconvolution methods offer a computationally tractable alternative by inferring cell-type proportions from bulk data using single-cell reference profiles. However, existing approaches struggle to distinguish closely related cell types with similar expression patterns.

DeCovarT addresses this limitation by incorporating gene regulatory networks (GRNs) into a generative model. The approach assumes that cell-type-specific expression profiles follow multivariate Gaussian distributions whose covariance structure is estimated via Graphical Lasso, with optional integration of mechanistic prior knowledge as regularisation weights. Bulk expression is then modelled as a weighted convolution of these cell-type distributions, identifiable to a multivariate Gaussian which naturally accommodates biological variability within each cell type. Crucially, using an explicit generative framework enables the computation of asymptotic confidence intervals of the estimated parameters, providing rigorous uncertainty quantification of cell-type composition variations.

Benchmarking on simulated data demonstrates that DeCovarT consistently outperforms alternative approaches, particularly when cell-type expression profiles substantially overlap.

As a proof-of-concept application, DeCovarT is applied to gastruloids, 3D stem-cell-derived models of early embryonic development, where bulk transcriptomics is routinely used to monitor differentiation over time. From the inferred cell-type compositions, a time-dependent organoid potency score is derived, quantifying the developmental potential of the organoid at each sampled time point and enabling dynamic tracking of gastruloid maturation trajectories.

URL

https://cnrs.hal.science/hal-04208010

https://github.com/bastienchassagnol/DeCovarT

Defining Populations in the Presence of Admixture: Insights from Saccharomyces cerevisiae Genomics

Poster

Louis OLLIVIER [1], ***Fanny Pouyet*** [2], ***Gilles Fischer*** [3]

1. *Université Paris-Saclay, CNRS, Laboratoire Interdisciplinaire des Sciences du Numérique,* ***2.*** *LISN, Université Paris-Saclay,* ***3.*** *Sorbonne Université, CNRS, CQSB*

Abstract

Understanding how genetic variation is structured within and among populations is central to inferring evolutionary history and population genomics. In many species, population boundaries are blurred by ongoing gene flow and admixture, complicating the identification of biologically meaningful groups. The budding yeast Saccharomyces cerevisiae provides an ideal system to investigate these challenges, as it combines extensive ecological diversity, a long history of domestication, and abundant genomic resources. Large-scale sequencing projects have revealed deeply diverged wild lineages, multiple domesticated clades, and widespread mosaic genomes, highlighting both the progress and the increasing complexity of yeast population structure.

Here, we develop an integrated framework to define robust population units and characterize admixture in 3,570 S. cerevisiae strains. Whereas previous studies have examined admixture in isolated case studies, our framework enables a comprehensive, species-wide analysis of admixture patterns, allowing comparisons between admixed and unadmixed samples and facilitating the interpretation of population-level patterns. Using ensembles of clustering analyses, phylogenetic and population genetic validation, f-statistics, and local-ancestry inference, we identified conservative sets of 38 unadmixed populations (~1,600 isolates) and identified a clear admixture signal in approximately 1,300 isolates clustered into 69 lineages sharing common ancestries. We show that admixed genomes display similar heterozygosity but less polyploidy and heterothallism than unadmixed populations. However, this pattern changes when ecological origins are taken into account, suggesting that reproductive dynamics vary across ecological contexts. Together, our analyses provide a reproducible framework for studying population structure and admixture in yeast and establish a basis for future demographic and genotype-phenotype investigations in important domesticated species.

Designing genome annotation tools to investigate the evolution of bioenergetic enzymes

Poster

Alexis Nguyen [1], ***Sophie Abby*** [1], ***Fabien Pierrel*** [1], ***Barbara Schoepp-Cothenet*** [2], ***Frauke Baymann*** [2], ***Axel Magalon*** [3], ***Gwendoline Degré*** [2]

1. *Univ. Grenoble Alpes, CNRS, UMR 5525, VetAgro Sup, Grenoble INP, TIMC, 38000 Grenoble, France,* ***2.*** *Aix-Marseille Université, CNRS, BIP-UMR 7281, Marseille, France,* ***3.*** *Aix Marseille Université, CNRS, Laboratoire de Chimie Bactérienne (UMR7283), IMM, IM2B, 13402 Marseille, France*

Abstract

Quinones are lipophilic molecules that carry electrons between enzymes of photosynthetic and respiratory chains. They are thus essential for the generation of the proton gradient that drives ATP synthase and the generation of ATP, the energy currency of the cell.

We aim at investigating how enzymes from photosynthetic and respiratory chains have adapted their quinone-binding sites (Q-sites) to one or more types of quinones over time.

The project involves designing a bioinformatics pipeline for genome annotation using MacSyFinder, a bioinformatics tool for genome annotation of macromolecular systems and biochemical pathways, which utilizes sequence similarity search and co-localization rules of the genes involved in the biological function to annotate, that are typical of bacteria and archaea.

In order to design MacSyFinder models, we used curated sets of reference sequences for the generation of HMM (hidden Markov Model) protein profiles, and a priori biological knowledge on the genomic organization of the genes to annotate. This allowed to develop models for the different quinone synthesis pathways ,and quinone-interacting bioenergetic enzymes on the other hand. models are currently being developed for 5 different Q pathways, including a total of 95 HMM protein profiles

In order to test the consistency of the results obtained on a large dataset, the pipeline has been benchmarked by comparing its results with those from a recent article from the team (total of 27,092 genomes). Few differences were observed, and those were consistent with the existing literature.

The next step will involve integrating quinone annotation with Q-enzymes annotation by mapping them onto a species tree, helping identify evolutionary changes of Q-enzymes related to quinone repertoire variations. This will allow us to investigate possible correlations between quinone types and the structure of Q-enzymes quinone-binding sites, a first step towards the understanding of Q-enzymes adaptation to quinone repertoire along evolution.

Development in R of a processing pipeline integrated into an interface for flow cytometry data analysis

Poster

Camellia Lambert [1]

1. *Institut Cochin, Inserm U1016, CNRS UMR8104, Plateformes CYBIO et BIOINFORMAT'IC, Université Paris Cité, Paris, France*

Abstract

Flow cytometry enables individual analysis of thousands of cells by measuring fluorescence after laser excitation. With technological advances and increasing marker numbers, datasets have considerably expanded, making analysis challenging. Research platforms rely on expensive and complex software, reducing access to advanced analyses. Within BIOINFORMAT'IC and CYBIO platforms from Institut Cochin, this project develops an open-source R-based pipeline. Following literature standards, this tool will be integrated into a user-friendly interface for biologists.

Directly from the cytometer, raw FCS files are integrated into a workflow developed via the Bioconductor ecosystem. The analysis begins with a fluorescence spillover correction step (compensation or unmixing) through the calculation of a matrix. Quality control is then performed using the PeacoQC or flowAI algorithms to eliminate acquisition instabilities, and others functions to remove debris, doublets, and dead cells. Then an arcsinh or logicle transformation is applied to facilitate population gating. At the end of this process, files are ready for high-dimensional analyses (clustering, UMAP, etc…). Finally, this pipeline is integrated into a R-Shiny interface and will be deployed on the Institut Cochin's internal database.

Compensation and preprocessing phases were tested on a 4-color dataset to assess code reliability. Comparisons with OMIQ software show convergent results. The calculated spillover matrix is identical, thus validating the accuracy of the calculation. Similarly, debris and doublet removal procedures produce comparable cell populations. The creation of RDS objects at each stage ensures full analysis of traceability and offers flexibility for the user. Finally, functions were designed to apply uniform settings for consistency, while allowing manual adjustments for specific samples if necessary.

The immediate perspective is its full integration into the Shiny interface for broader accessibility. This project demonstrates that a free alternative is viable, offering laboratories greater autonomy from commercial solutions without compromising scientific rigor.

Development of a metabolic score predictive of survival in patients with Multiple Myeloma

Poster

Philippe Laurent [1], ***Alizée Steer*** [1], ***Elina Alaterre*** [1], ***Angélique Bruyer*** [1], ***Jérôme Moreaux*** [2]

1. *Diag2Tec,* ***2.*** *IGH*

Abstract

Background: Metabolic activity plays an important role in treatment response across various cancers. However, in Multiple Myeloma (MM), no gene-expression-based score integrating metabolic information has yet been reported to predict patient survival. Because metabolic measurements are rarely available in patient cohorts, we used functional metabolic profiling from Human Myeloma Cell Lines (HMCL), representative of the molecular heterogeneity of the MM disease[1], together with RNA sequencing data to develop a gene-expression–based metabolic score.

Method: Eight glycolytic or oxidative phosphorylation parameters, measured or calculated using the Seahorse XF Mito Stress test Kit, were combined with the raw counts data of 105 metabolic genes from 26 HMCLs. After raw count normalization (RLE method)[2], log2-transformation and standardization, we developed predictive scores derived from machine learning approaches including Elastic Net regression[3,4], partial least squares (PLS) regression, and stepwise optimization. Correlations between predictive scores and metabolic measurements were then evaluated, along with their prognostic value, using the maxstat method[5] in both Montpellier[6] (n=102) and CoMMpass[7] (n=669) cohorts.

Results: Geometric mean gene expression levels were highly correlated across datasets (Spearman r = 0.85–0.95, p < 0.001), supporting cross-cohort comparability. The best predictive score was derived from the "Glycolysis ATP Production Rate" calculated value, which consists of a linear combination of 11 metabolic genes (6 glycolytic and 5 oxidative phosphorylation genes). This predictive score was significantly correlated with 7 of the 8 functional metabolic parameters (r = 0.42-0.93) and significantly stratified MM patients in the Montpellier (low-risk = 91, high-risk = 11, p = 0.02) and CoMMpass (low-risk = 474, high-risk = 195, p = 0.001) cohorts.

Conclusions: We developed an 11-gene expression-based score associated with high glycolytic metabolism in HMCL and identified high-risk MM patients, suggesting increased glycolytic activity in this subgroup.

URL

- DESeq2. *Bioconductor* http://bioconductor.org/packages/DESeq2/
- Elastic Net Regression | NRT Online Library for Data Science and Human Behavior. https://nrt-library.github.io/nrt-library/elastic_nets.html

Developping a reusable and robust microbiota analysis pipeline using non-robust methods

Poster

Corentin LUCAS [1], ***Emmanuelle BECKER*** [1], ***Yann Le Cunff*** [1]

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

Abstract

Differential abundance analysis is a classic method in microbiome research, but the available tools are often misused or applied interchangeably. This leads to inconsistent and unreliable results, which is problematic when studying human health or other critical topics.

In this study, we developed a robust and reproducible pipeline for differential abundance analysis, applied to a cohort of 501 Crohn's disease samples composed of 500 amplicon sequence variants (ASVs), that were preprocessed and filtered to yield 197 ASVs. Our approach uses several of the largely used tools to combine the results and then returns a robust microbiome signature with a measure of robustness and consistency ofr the methods on the dataset. The goal is to try to guide current inconsistent practices toward a more rational and reliable workflow, producing results that are both accurate and biologically meaningful.

Beyond this specific dataset, this pipeline is designed to be reusable and adaptable, providing a solid foundation for differential abundance analysis in other microbiome-related studies.

Digital Twins of Organoids: a Knowledge Graph of human organoids omics dataset

Poster

Kenza Zeghari [1], ***Youssef Boulaimen*** [1], ***Bastien Chassagnol*** [1], ***Marielle Péré*** [1], ***Anaïs Baudot*** [2]

1. *Aix-Marseille Université,* ***2.*** *CNRS*

Abstract

Human organoids emerged as pivotal models for studying organ and tissues, model diseases, or discover novel therapeutics. This surge in organoid research has been accompanied by the production of large volumes of heterogeneous omics datasets. To enable cross-study integration and complex semantic queries (e.g., identifying all datasets associated with specific perturbagens, disease-related genes, or accessing dataset using specific sequencing techniques across diverse organoid types), there is a critical need for a structured, interconnected knowledge representation, i.e. a Knowledge Graph (KG).

We designed an automated workflow based on Large Language Models (LLM) to extract and standardize metadata from summary files associated with human organoids omics datasets archived in public repositories. We used Retrieval-Augmented Generation (RAG) to supply the LLM with relevant context. To ensure accuracy and standardized annotations, the pipeline integrates real-time access to the EBI Ontology Lookup Service (OLS), enabling retrieval of ontology-compliant values and reducing hallucination risks. Output fidelity is ensured through a triple-redundancy execution model, in which final annotations require majority consensus across independent runs. The pipeline was validated against a gold-standard corpus of 50 manually annotated files, achieving 88% extraction accuracy. This high-precision approach will allow us to consolidate 993 datasets into a KG.

This resource offers a scalable foundation for the organoid community, supporting not only efficient data reuse but also downstream meta-analyses and biomarker identification across diverse organoid systems and perturbagen conditions.

DynAA: Characterizing the dynamics of antibody-antigen interfaces using Molecular Dynamics simulations

Poster

Louise LAM [1], ***Ayşe Berçin Barlas*** [2], ***Ezgi Karaca*** [3], ***Alessandro Masiero*** [4], ***Catherine Prades*** [5], ***Chantal Prévost*** [6], ***Sophie Sacquin-Mora*** [6]

1. *Laboratoire de Biochimie Théorique, Institut de Biologie Physico-Chimique, CNRS UMR8266, Université Paris Cité, Paris, France - Sanofi R&D, Computational Biology, Vitry-sur-Seine, France,* ***2.*** *Izmir International Biomedicine and Genome Institute, Dokuz Eylul University, Izmir, Türkiye - Computational Structural Biology Laboratory, Izmir Biomedicine and Genome Center, Izmir, Türkiye,* ***3.*** *Izmir International Biomedicine and Genome Institute, Dokuz Eylul University, Izmir, Türkiye,* ***4.*** *Sanofi R&D, Computational Biology, Vitry-sur-Seine, France,* ***5.*** *Sanofi R&D, Computational Biology, Vitry-sur-Seine, France,,* ***6.*** *IBPC, LBT, UMR 8266, CNRS, UPCité*

Abstract

Studying the antibody-antigen (Ab-Ag) interfaces is crucial for understanding the molecular basis of antibody recognition and for designing effective therapeutics. Current studies mainly focus on static structures coming from experimental methods such as X-ray crystallography, which do not capture the dynamic nature of these interfaces known to present specific binding behavior. This is why, in this work, we used trajectories from all-atom classical Molecular Dynamics (MD) simulations to add a temporal dimension to the analysis of Ab-Ag interfaces. We performed simulations on 212 non-redundant protein-protein complexes from the well-known open-source Docking Benchmark 5.5 dataset [1], including 57 Ab-Ag complexes classified as AA (Ag-Double chain Ab) or AS (Ag-Single chain Ab). These runs were made possible by an exceptional grant of 90 million CPU hours from GENCI, allowing us to generate at least 3 replicates of 100 ns-long trajectories for each complex. All simulation data, now accessible through DynaRepo [2] and DynaBench [3] repositories, were analyzed using DynaPIN [4], a new analysis pipeline that combines structural stability metrics, interface quality assessment, residue-level energetics, and interaction network analysis to comprehensively characterize protein-protein interaction (PPI) dynamics. To our knowledge, this work will constitute the first large-scale analysis comparing Ab-Ag interface dynamics to other PPI types, with future extensions focusing specifically on AA versus AS interaction mechanisms. The results are expected to reveal dynamic signatures specific to Ab-Ag interfaces and pave the way for the identification of key residues/structures critical for AA and AS assembly stability, with important implications for therapeutic antibody design targeting specific antigens.

Effect of ultra-processed food consumption on the human sperm epigenome

Poster

ELZA BERSANOUKAEVA [1], ***Marie-Charlotte Dumargne*** [1]

1. *IPMC*

Abstract

Ultra-processed foods (UPFs; NOVA category 4) now account for more than half of daily energy intake in many Western countries and have been consistently associated with adverse metabolic and cardiovascular outcomes. These foods contain industrial additives, contaminants, and packaging-derived chemicals, and higher consumption has been linked to increased urinary concentrations of phthalates and bisphenols. Our group and others showed that preconceptional consumption of certain diets and pollutants alter the metabolic and behavioral phenotype of the offspring, through gametic epigenetic inheritance.

Here we hypothesized that consumption of ultra-processed foods influences not only metabolic and reproductive outcomes, but the DNA methylation signature of spermatozoa. To test this, we use monozygotic twin pairs, providing a highly controlled design that minimizes genetic confounding while isolating environmental effects. Five monozygotic twin pairs followed either an ultra-processed or unprocessed diet. Caloric intake was held constant across conditions to isolate the effects of food processing for three weeks. Sperm was collected and analyzed using Oxford Nanopore long-read sequencing to profile DNA methylation (5mC), hydroxymethylation (5hmC), and genomic variation.

Twins consuming the ultra-processed diet exhibited increased body weight, altered LDL:HDL ratios, and a trend toward reduced sperm motility. Blood analyses also suggest metabolic and inflammatory changes associated with the dietary intervention, including alterations in mineral levels such as selenium and hematological parameters such as platelet counts.

Preliminary analyses indicate a global decrease in sperm DNA methylation following the ultra-processed diet compared with the unprocessed diet, along with chromosome-specific differential methylation patterns enriched at CpG islands within promoters and intronic regions. Ongoing analyses are currently testing different statistical thresholds and analytical conditions to ensure robustness. In addition, we are analyzing sperm DNA to identify potential mutations or transposable elements insertions associated with dietary exposure. All analyses are ongoing, and full results will be presented.

eHGTDB: A web platform for the exploration and visualization of horizontal gene transfer events in eukaryotes

Poster

Corinne Rancurel [1], ***Mathéo Coiffet*** [2], ***Arthur Péré*** [2], ***Dominique Colinet*** [2], ***Etienne G.J. Danchin*** [2]

1. *PHYBAC (EMR7006), CNRS, INRAE, Université Côte d'Azur, 400 route des chappes, 06903, Sophia Antipolis, France,* ***2.*** *Institut Sophia Agrobiotech (UMR1355), INRAE, Université Côte d'Azur, 400 route des chappes, 06903, Sophia Antipolis, France*

Abstract

Horizontal gene transfer (HGT) is a major evolutionary mechanism enabling the acquisition of genetic material from unrelated organisms. While extensively documented in prokaryotes, growing evidence indicates that HGT also occurs in eukaryotes, where it can contribute to metabolic innovation, ecological adaptation, and host–pathogen interactions. Several large-scale studies have identified candidate HGT events across diverse eukaryotic lineages using comparative genomics, phylogenetic analyses, and sequence similarity searches. However, these datasets are often dispersed across publications and stored in heterogeneous formats, which limits their accessibility and comparative exploration.

To address this limitation, we developed eHGTDB, a database dedicated to the collection, exploration, and visualization of predicted horizontal gene transfer events in eukaryotic genomes. The platform integrates predictions generated using the Alienness (**1**) and AvP (**2**) pipelines and aims to provide a unified environment for exploring candidate transferred genes together with their genomic and evolutionary contexts.

The platform is implemented as a modular and containerized web application orchestrated with Docker Compose to ensure reproducibility and portability. The backend relies on the Django framework coupled with a MariaDB relational database for efficient storage and querying of genomic annotations and HGT prediction metadata. The frontend is built using modern web technologies including Node.js and Vue.js, enabling a responsive and interactive user interface. Data exploration is supported by dynamic visualizations implemented with D3.js and circular genome representations based on Circos, allowing users to investigate gene distributions, genomic contexts, and potential transfer events.

By centralizing HGT predictions and providing interactive exploration tools, eHGTDB facilitates comparative analyses across eukaryotic taxa and supports research on genome evolution, functional innovation, and host–pathogen interactions.

1. Rancurel et al., 2017. *Genes*. doi:10.3390/genes8100248. https://alienness.sophia.inrae.fr
2. Koutsovoulos et al., 2022. *PLOS Computational Biology*. doi:10.1371/journal.pcbi.1010686. https://github.com/GDKO/AvP

Elucidation and modeling of the insertion mechanism driving high pathogenicity avian influenza emergence.

Poster

Aldair Martin Martinez Pineda [1], ***Bertille Pouget*** [2], ***Gabriel Dupré*** [2], ***Claire Hoede*** [1], ***Christine Gaspin*** [3], ***Romain Volmer*** [2]

1. *Université de Toulouse, INRAE, UR 875 MIAT, F-31320, Castanet-Tolosan, France,* ***2.*** *Ecole Nationale Vétérinaire de Toulouse, Université de Toulouse, ENVT, INRAE, IHAP, UMR 1225, Toulouse, France,* ***3.*** *Université de Toulouse, INRAE, UR 875 MIAT, F-31320*

Abstract

Highly pathogenic avian influenza viruses (HPAIV) evolve from low-pathogenic avian influenza viruses (LPAIV) of H5 and H7 subtypes through nucleotide insertions that introduce a multibasic cleavage site (MBCS) in the hemagglutinin (HA). By combining experimental evolution and modeling approaches, we show that the product–template dimer interaction formed within the catalytic site of the viral polymerase during RNA replication constitutes the main driver of nucleotide insertions in H5 viruses and in the majority of H7 viruses.

In particular, we demonstrate that specific adenine-rich sequences in the cRNA orientation promote dissociation of the product–template dimer and upstream rehybridization, allowing the viral polymerase to backtrack and duplicate nucleotide sequences. Using this mechanistic insight, we developed a mathematical model based on the thermodynamic stability of the product–template dimer that predicts the risk of MBCS acquisition through nucleotide insertions.

Applying this predictive framework to H5 and H7 sequences available in the GISAID database, we identified sequences with a high probability of acquiring insertions at the HA cleavage site. We also identified specific sequence backgrounds associated with historical HPAIV emergence events. This sequence-based risk assessment provides a potential early-warning strategy and supports source-level control of viruses at risk of evolving toward highly pathogenic forms.

Enhancing Genomic Prediction Accuracy for Complex Traits in Cassava (Manihot esculenta) Through Pangenome-Informed Variant Calling

Poster

Isaac ABEGUNDE [1], ***Olabode Onile-ere*** [1], ***Fidele Tiendrebeogo*** [2], ***Justin S. PITA*** [2], ***Emmanuel Idehen*** [3], ***Angela ENI*** [2]

1. *Central and West African Virus Epidemiology Program, Covenant University Hub, Km. 10 Idiroko Road, Canaan Land, Ota, Ogun State, Nigeria,* ***2.*** *Regional Center of Excellence for Transboundary Plant Pathogens (Central and West African Virus Epidemiology, WAVE), Université Felix Houphouët-Boigny (UFHB), 01 BPV 34 Abidjan 01, Côte d'Ivoire,* ***3.*** *Department of Plant Breeding and Seed Technology, Federal University of Agriculture, Abeokuta, Ogun state, Nigeria.*

Abstract

Cassava (*Manihot esculenta*) is a critical food security crop, yet its highly heterozygous genome challenges traditional genomic selection (GS) models that rely on single linear references. Current single-nucleotide polymorphism (SNP)-based methods often miss complex structural variations (SVs), which significantly influence complex agronomic traits and limit prediction accuracy. This study aims to evaluate whether incorporating graph-based pangenomes and SVs can enhance genomic prediction models. We developed a pangenome-informed pipeline to capture the full spectrum of genetic diversity. A graph-based pangenome was constructed using long-read sequences (PacBio/Nanopore) from a diverse panel of 30 accessions, including wild relatives and improved lines. Next, short-read data from a breeding population was mapped to the graph using VG Giraffe, followed by graph-based variant calling to simultaneously genotype SNPs and SVs to construct a comprehensive pangenomic relationship matrix (Pangenome-G). Analyses are currently underway. We hypothesize that incorporating graph-captured SVs will yield a measurable increase in genomic prediction accuracy (ρ) across target traits compared to standard SNP-based Reference-GBLUP models. This improvement is expected to be most pronounced for highly polygenic yield components and complex quality traits. Establishing this pangenome-informed approach would provide a robust framework for genomic selection in heterozygous crops, leveraging structural diversity to accelerate genetic gain in cassava breeding.

URL

https://github.com/isaaco25/Cassava_Pangenome_Build

Evaluating protein representations from domain architectures

Poster

Sheyenne NGUYEN [1], ***Philippe Ortet*** [1], ***Louison Silly*** [1]

1. *BIAM - CEA Cadarache*

Abstract

Artificial intelligence approaches applied to proteins predominantly rely on the analysis of amino acid sequences like the large-scale protein language model ProtBERT, trained directly on sequences [1]. These approaches, when applied to long sequences (>3000 aa) can be computationally costly. Recently, new studies led to the development of models based on more compact representations, such as domain architectures, exemplified by dom2vec [2]. Domain architectures can be defined by InterPro as the linear arrangement of conserved protein domains along a sequence, reflecting its structural and functional organization.

However, existing evaluation frameworks, such as FLIP [3] and TAPE [4], are designed and parameterized for sequence-centered models and do not allow for an appropriate assessment of domain-based approaches.

We therefore propose a benchmark specifically dedicated to domain architectures, providing a standardized, reliable, and comparative framework (Figure 1). Models are trained and evaluated on a set of tasks covering different levels of biological questions, from protein functional characterization to the study of their interactions and regulatory mechanisms. These tasks assess the ability of models to capture biological information from domain architectures, in order to test the hypothesis that these representations can be used to predict various protein properties.

This new framework, filling an important gap in current evaluation strategies, aims to evaluate the informational relevance of domains and to identify more compact and interpretable representations for the classification of complex biological systems.

[1]Elnaggar, Ahmed, et al. “ProtTrans: Towards cracking the language of life’s code through self-supervised learning.” (2020).

[2]Melidis, Damianos P., and Wolfgang Nejdl. “Capturing protein domain structure and function using self-supervision on domain architectures.” Algorithms 14.1 (2021): 28.

[3]Dallago, Christian, et al. “FLIP: Benchmark tasks in fitness landscape inference for proteins.” bioRxiv (2021): 2021-11.

[4]Rao, Roshan, et al. “Evaluating protein transfer learning with TAPE.” Advances in neural information processing systems 32 (2019).

Evaluation of 7 jDR methods for multi-omics survival prediction: a benchmark study on 18 cancer datasets

Poster

Vincent Le Goff [1], ***Vincent Guillemot*** [2], ***Cathy Philippe*** [3], ***Gwendoline Mendes*** [4], ***Jean-François Deleuze*** [5], ***Edith Le Floch*** [1], ***Arnaud Gloaguen*** [1]

1. *Mathématiques et Statistiques, CNRGH, Institut de Biologie François-Jacob, CEA,* ***2.*** *Institut Pasteur, Université Paris Cité, Bioinformatics and Biostatistics Hub,* ***3.*** *Neurospin, CEA, Université Paris-Saclay,* ***4.*** *CentraleSupélec,* ***5.*** *Centre National de Recherche en Génomique Humaine (CNRGH), IBFJ, CEA, Université de Paris-Saclay, Evry, France*

Abstract

The complexity of certain diseases requires multiple measurements on the human genome to fully understand the underlying dysregulations, which led to the generation of multi-omics datasets. The high-dimensionality and heterogeneity inherent to multi-omics datasets appear to be challenging to analyse. Nevertheless, in the context of survival analysis and exploiting 18 distinct cancer datasets from TCGA (Herrmann et al., 2021) showed that models considering the inherent group structure of multi-omics datasets can help leverage their full potential. Yet, from their conclusions, it remains unclear whether the joint analysis of molecular and clinical data increases predictive power in comparison to the analysis of clinical data only.

Extending this benchmark, we aim to tackle this limitation by exploring methods that extract links between omics data blocks by employing 7 joint Dimension Reduction (jDR) techniques: RGCCA, JIVE, IntNMF, MCIA, iCluster, MOFA, tICA. For each method, several combinations of hyperparameters are tested. Our approach initially estimates a reduced space from molecular data alone via unsupervised or supervised techniques, followed by survival prediction using a Cox model on this joint reduced space, with and without clinical data. To be able to properly benchmark these methods, we had to develop appropriate prediction methods for IntNMF, MCIA and iCluster. We demonstrate that jDR methods, when combined with a Cox model, can significantly outperform traditional ones. When jointly analyzing clinical and omics data, some of the jDR methods perform significantly better than the baseline model (Cox model on clinical data only). Furthermore, when prediction is done only from omics data, most of the compared methods perform significantly lower than this baseline, yet some methods are able to match, and on some cancers exceed, the performance of the baseline. This can be interesting in an exploratory setting, e.g. to identify key genes or pathways impacting patients' survival.

Evaluation of Helixer for structural genome annotation in non-model organisms

Poster

Audrey Onfroy [1], ***Sophie Lemoine*** [1], ***Catherine Senamaud-Beaufort*** [1], ***Laurent Jourdren*** [1], ***Morgane Thomas-Chollier*** [1]

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

Abstract

Context

Accurate structural genome annotation is essential for reliable transcript quantification in transcriptomics analyses, yet high-quality annotations usually require extensive experimental data that are often unavailable for non-model organisms. Existing annotation pipelines can be difficult to interpret and tune. Recently, tools using pre-trained deep learning models, such as Helixer [1], have been proposed to predict gene structures directly from genome assemblies. Here, we evaluated the reliability of such approaches in practical analysis workflows.

Results

A pipeline was developed using Nextflow and Docker to run Helixer from a genome assembly (FASTA) and generate predicted annotations (GFF) together with evaluation metrics. On the model plant Arabidopsis thaliana, Helixer showed high precision, indicating that most predicted transcripts correspond to known annotations. However, the tool shows lower sensitivity, with some loci missing.

We further evaluated Helixer on a non-model invertebrate by comparing predicted annotations with a reference annotation and a long-read RNA-seq–based annotation. Helixer predicted slightly longer transcripts but detected fewer genes. In single-cell RNA-seq data from the same organism, the default annotation resulted in ~70% of reads confidently mapped to exonic regions, compared with ~80% using the long-read annotation, while Helixer-based annotations resulted in <60%.

A suited annotation of the 3'UTR regions is essential to recover the signal from 10X-based scRNA-Seq data. Because Helixer predicts full transcript structures, we tested whether combining Helixer and long-read annotations could improve mapping. Although merging annotations is challenging, the consensus annotation provided only limited improvement.

Prospects

For the studied case, Helixer outputs annotations less sensitive than the reference or long-read–based annotations. Ongoing work focuses on species-specific retraining to refine its potential in annotating non-model species.

Acknowledgments

The GenomiqueENS core facility was supported by the France Génomique national infrastructure, funded as part of the "Investissements d'Avenir" program managed by the Agence Nationale de la Recherche (contract ANR-10-INBS-0009).

URL

References

[1] Holst F. et al., Helixer: ab initio prediction of primary eukaryotic gene models combining deep learning and a hidden Markov model , Nature Methods (2025). DOI: 10.1038/s41592-025-02939-1

Extending the Semantic Metabolomics Data Lake: Integrating Plant and Food Transformation Ontologies for Enhanced Knowledge Graphs

Poster

Isaac Karaman [1], Guillaume Laisney [2], Clement Frainay [2], Franck Giacomoni [3], Olivier Filangi [1], Magalie Weber [4]

1. *Institute for Genetics, Environment and Plant Protection (IGEPP), National Research Institute for Agriculture, Food and Environment (INRAE), Institut Agro, Université Rennes,* ***2.*** *Toxalim (Research Center in Food Toxicology), Université de Toulouse, INRAE, ENVT, INP-Purpan, UPS,* ***3.*** *University Clermont Auvergne, INRAE, UNH, Metabolism Exploration Platform, MetaboHUB Clermont,* ***4.*** *INRAE, UR BIA (Biopolymers Interactions Assemblies)*

Abstract

The “Semantic Metabolomics Data Lake” is a Big Data infrastructure that generates and leverages knowledge graphs to contextualize data from metabolomics platforms. It relies on automated workflows that reliably and regularly ingest RDF data from from specialized bioinformatics databases, primarily sourced from PubMed and focused on the biomedical domain. The objective of this work was to adapt this workflow to produce additional graphs covering plant science and food transformation processes.

Scientific literature available in repositories complementary to PubMed—such as ISTEX, PMC, and its European counterpart EuropePMC—abounds with information on the key role of metabolites in plant products. These metabolites influence disease resistance, environmental interactions, and organoleptic qualities, and are essential for understanding plant characteristics, stress responses, and transformation potential. Reference ontologies including the Plant Ontology, Trait Ontology, and Plant Experimental Conditions Ontology (Planteome project), complemented by INRAE’s TransformON for bioresource transformation processes, provide structured semantic standards. These ontologies play a crucial role in text annotation of scientific literature and are essential for generating semantically rich knowledge graphs.

A preliminary study assessed the value of the targeted sources through quantitative and qualitative coverage comparisons. An initial analysis examined the number of articles on plant phytochemistry and food chemistry to identify overlaps and complementarities. Additionally, a qualitative evaluation of publication years and journal publishers was conducted. This study confirmed the relevance of integrating these sources into the “Semantic Metabolomics Data Lake.” The integration was achieved by developing dedicated Python modules within Airflow, the infrastructure’s workflow management tool.

This work enabled the implementation of the FORVM methodology, which explores graphs linking metabolites to other concepts from ontologies of interest. This approach addresses key scientific questions in plant science, such as the impact of processing on polyphenol reactivity, responses to water stress, or fruit ripening.

URL

https://hal.inrae.fr/hal-05075616v1

Family-level classification of viral contigs using deep learning

Poster

Emma Soufir [1], ***Florian CHARRIAT*** [1], ***Antoni Exbrayat*** [1], ***Ilka Engelmann*** [2], ***Maximilien Servajean*** [3], ***Serafin Gutierrez*** [1]

1. *ASTRE, CIRAD,* ***2.*** *CHU Montpellier,* ***3.*** *ADVANSE, LIRMM*

Abstract

Metagenomics applied to viruses has greatly improved the ability to study virus diversity. However, this approach has also come with its own challenges. Among others, sequence assemblies often result in fragmented genomes which complicate taxonomic identification using traditional methods. Indeed, most taxonomic classification methods rely heavily on sequence similarities against reference databases or marker gene detection, and therefore depend on the completeness and quality of these databases. The limited coverage of viral diversity in these databases often leads to inaccurate or unresolved annotations. At the same time, recent advances in deep learning, particularly the use of large language models with genomic sequences, offer promising solutions for sequence-based classification tasks.

Here, we present a method for family-level classification of viral contigs using a fine-tuned DNABERT-2 model. For training, we generated a curated viral sequence database. We used the ICTV MSL to define the taxonomy and focused specifically on eukaryotic infecting viruses. Additional sequences were also retrieved to expand the dataset. Genomes were cut into 1 000 b fragments to mimic contig-like sequences. To better reflect real scenarios, we investigated several training strategies designed to simulate the presence of previously unseen species, thereby evaluating the model's ability to generalize beyond known taxa. We also explored different training optimizations, like data augmentation, to enhance predictive performance and robustness. We provide a benchmark comparing multiple methods, like Kraken, evaluated on simulated and real metagenomic datasets. Our results highlight the potential of transformer-based models for accurate viral taxonomic classification in metagenomic studies. We also analyze the errors made by the model to better understand its limitations and strengths in comparison with traditional methods.

Few-shot learning strategy for Predicting Meropenem Resistance genes in Escherichia coli

Poster

Meriem YOUSSEF [1], ***David VALLENET*** [2], ***Alexandra CALTEAU*** [2], ***Guillaume GAUTREAU*** [3]

1. *LABGeM, Génomique Métabolique, CEA, Genoscope, Institut François Jacob, Université d'Évry, Université Paris-Saclay, CNRS,,* ***2.*** *LABGeM, Génomique Métabolique, Genoscope, Institut François Jacob, CEA, CNRS, Université Évry, Université Paris-Saclay,* ***3.*** *Université Paris-Saclay, INRAE, MaIAGE, 78350, Jouy-en-Josas, France*

Abstract

Trained on large-scale data, foundation models have shown potential for few-shot learning [1], enabling predictive models to generalize from limited labeled datasets through transfer learning. Predicting uncatalogued genomic features from limited labeled genomic data remains a challenging machine-learning problem due to the complex structure of bacterial genomes and the variable number of genes across isolates. To address these challenges, we propose harnessing the potential of genomic LLMs (genomic Language Models) to classify meropenem resistance in *Escherichia coli.*

Our approach leverages *Bacformer* [2], an encoder-based Transformer model that generates genome-scale, contextualized protein embeddings. Whole-genome assemblies of *E. coli* labeled as sensitive or resistant isolates were collected from the BV-BRC database [3] and then annotated to identify protein-coding genes. Unlike methods that encode proteins independently, *Bacformer* processes proteins in their genomic order, enabling each embedding to capture both local and long-range dependencies within the genome. These contextualized embeddings are then used as input features for a resistance prediction model based on a gated attention-based multiple-instance learning mechanism [4]. This design is expected to improve both predictive performance and interpretability by revealing which proteins contribute most strongly to the model's predictions.

Although antibiotic resistance genes are often already well referenced in databases that enable direct annotation, making them accessible without machine learning, they provide a useful benchmark for validating this generic and interpretable predictive approach. This strategy could more broadly support the prediction of any genomic features linked to a genome label in few-shot settings offering an alternative to bacterial GWAS with predictions that can be assessed here against known annotations.

FiFi: Functional Inference from Fungal ITS, A bioinformatics tool to infer fungal metagenomes from ITS data

Poster

Maëlle Pomiès [1], ***Marc Buée*** [1], ***Lucas Auer*** [1]

1. *Université de Lorraine, INRAE, UMR 1136 Interactions Arbres/Microorganismes, 54280, Champenoux, France*

Abstract

Background

Fungi play a key role in ecosystem functioning, particularly in biogeochemical cycles, organic matter decomposition, and plant–microorganism interactions. Fungal communities are commonly studied using metabarcoding approaches targeting ribosomal ITS (Internal Transcribed Spacer), the standard marker for fungal taxonomic identification.

However, ITS metabarcoding mainly provides taxonomic information and offers limited insight into the ecological or metabolic functions of detected organisms. In bacteria, tools such as PICRUSt2[1] can infer functional potential from metabarcoding data, but their phylogeny-based approach is not suitable for fungal ITS, which is not a reliable phylogenetic marker. Only a few methods exist for fungi, such as FunFun[2], but it infers gene content from a single ITS copy (whereas fungal genomes contain multiple ITS copies with substantial intra-genomic diversity), it does not infer metabolic pathways like PICRUSt2, and it relies on a limited genomic reference.

Method

To address these limitations, we constructed a large database of fungal genomic resources and developed a bioinformatics tool to infer the metagenome of fungal communities from their ITS sequences.

The reference database integrates 278,586 ITS sequences extracted from the 3,623 fungal genomes in MycoCosm and from UNITE[4] ITS database. ITS sequences can then be inferred using various k-mer methods (cosine distance, containment distance, or back_to_sequences[5]) or BLAST alignment. The closest references are selected using a k-nearest-neighbor (k-NN) strategy. The inferred metagenomes are then constructed based on similarity measures and corresponding genomes.

Conclusions and Perspectives

Evaluated on 20 sets of 1,000 known sequences, back_to_sequences is the best-performing distance, showing the highest percentage of correct taxonomic annotation. FiFi has been tested on fungal metabarcoding datasets from various ecosystems, ranging from forest soils to food-related systems, to ensure that its community coverage reaches levels sufficient to make the resulting functional inferences reliable.

Frhap: A flexible Snakemake Workflow for haplotype frequency estimation in tGBS data

Poster

Abdelkarim Wahnou [1], ***Aurélie Canaguier*** [1], ***Damien Hinsinger*** [1], ***Stéphane Nicolas*** [2], ***Raphaël Minguella*** [2], ***Patricia Faivre-Rampant*** [1]

1. *Université Paris-Saclay, INRAE, Centre Île-de-France Versailles-Saclay, EPGV,* ***2.*** *Université Paris-Saclay, INRAE, CNRS, AgroParisTech, GQE – Le Moulon*

Abstract

Next generation sequencing technologies enable large-scale Single Nucleotide Polymorphisms (SNPs) discovery. These SNPs, considered in combination, form haplotypes, which are widely used as markers for studying genetic diversity, population structure and trait association. Especially, microhaplotypes, i.e., closely linked SNPs within a short DNA fragment that are inherited together, appear promising in better describing fine-scale genomic diversity and in genome-wide association studies (GWAS). However, in pooled DNA samples, it can be challenging to identify microhaplotypes due to the mixture of sequences from multiple individuals.

Here we present Frhap, a Snakemake-based workflow that uses sequencing data to reconstruct microhaplotypes and infer their allele frequencies in pooled samples. It includes key processing steps, from reference genome preparation to generating a full summary report of the results. Frhap relies on established tools, including BWA-MEM for read alignment, offers the choice between GATK, FreeBayes, or VarDict for variant calling, followed by HARP for haplotype frequency estimation. Built to be as flexible as possible, Frhap is compatible with different wet-lab and sequencing configurations, including single-end or paired-end read data, with or without Unique Molecular Identifiers (UMI) information, amplicon clipping if applicable, and targeted-genotyping-by-sequencing (tGBS) datasets. Frhap is developed following the FAIR principles and will be openly available at https://forge.inrae.fr/epgv/.

We tested and validated Frhap using a tGBS dataset containing 2,400 samples and ~12,000 SNPs from the project MineLandDiv. This panel is mainly composed of maize landraces, chosen for their high genetic variability, making them well suited in order to demonstrate its ability to process large-scale sequencing data.

FROGS 5: A redesigned, modular pipeline for the comprehensive analysis of metabarcoding data.

Poster

Olivier Rué [1]***, Maria Bernard*** [2]***, Agoutin Gabryelle*** [3]***, Lucas Auer*** [4]***, Maëlle Pomiès*** [4]***, Géraldine Pascal*** [3]

1. *Université Paris-Saclay, INRAE, BioinfOmics, MIGALE bioinformatics facility, 78350, Jouy-en-Josas, France,* ***2.*** *Sigenae, GABI, INRAE, AgroParisTech, Université Paris-Saclay,* ***3.*** *GenPhySE, Université de Toulouse, INRAE, ENVT, 31326, Castanet Tolosan, France,* ***4.*** *Université de Lorraine, INRAE, UMR 1136 Interactions Arbres/Microorganismes, 54280, Champenoux, France*

Abstract

Background: FROGS1 is an open-source suite of tools dedicated to metabarcoding data analysis. It enables users to process amplicon sequencing data from raw reads to taxonomic and functional profiles, including statistical analyses, in a few steps. Designed to be accessible to both bioinformatics experts and non-specialists, FROGS can be used through the Galaxy2 platform or via the command line. Widely adopted by the community, the software has been downloaded nearly 40,000 times worldwide. To address evolving needs of metabarcoding analyses and new sequencing technologies, we introduce FROGS v.5, a version featuring a redesigned architecture and extended functionalities.

Results: FROGS v.5 introduces a modular organization structured into four complementary toolsets:

- FROGS_Core tools, including Core_Main and Core_Companion tools, which handle amplicon processing from raw reads to ASVs with taxonomic affiliations.
- FROGS_Stat tools, to statistical analyses of microbial community structure, diversity and differential composition.
- FROGS_Func tools, which infer the functional potential of microbial communities from ASVs.

A new read processing tool replaces the previous preprocessing and clustering steps, providing a unified workflow to generate ASVs. Users can now choose between the DADA23 denoising algorithm or the Swarm4 clustering approach.

FROGS also strengthens long-read metabarcoding analysis, introducing improved processing steps and a new chimera_denovo detection algorithm from VSEARCH5. In addition, FROGS now provides access to 135 reference databases, including in-house resources such as the GTDB 16S-ITS-23S database, designed for long-read analyses. These developments are supported by a new dedicated website (https://frogs.inrae.fr), featuring a modern visual identity, documentation and tutorials. The platform also highlights companion tools: Affiliation Explorer6, Easy16S7 and LEAP8.

Conclusions: With its redesigned architecture, expanded methodological options and improved support for long-read sequencing, FROGS v.5 provides a more flexible, complete and user-friendly solution for metabarcoding data analysis, facilitating reproducible workflows for expert and non-specialist.

URL

https://frogs.inrae.fr

From Waste to Enzymes: A Metagenomic Approach to Uncover Plastic-Degrading Microbes in Brazil

Poster

Julia Cantuti Gendre [1], ***Stéphanie Fouteau*** [2], ***Mark Stam*** [1], ***David Sanz Mata*** [3], ***Jorge Barriuso*** [3], ***Aleksandra Lazarova*** [3], ***Marli Camassola*** [4], ***Alicia Prieto*** [3], ***David VALLENET*** [5]

1. *LABGeM, Génomique Métabolique, CEA, Genoscope, Institut François Jacob, Université d'Évry, Université Paris-Saclay, CNRS,,* ***2.*** *LABGeM, Génomique Métabolique, CEA, Genoscope, Institut François Jacob, CNRS, Université d'Évry, Université Paris-Saclay, 91057 Evry, France,* ***3.*** *Centro de Investigaciones Biológicas Margarita Salas: Madrid, Madrid, Spain,* ***4.*** *Enzymes and Biomass Laboratory, Biotechnology Institute, University of Caxias Do Sul, Rua Francisco Getúlio Vargas, Caxias Do Sul, RS, 1130, 95070-560,,* ***5.*** *LABGeM, Génomique Métabolique, Genoscope, Institut François Jacob, CEA, CNRS, Université Évry, Université Paris-Saclay*

Abstract

The versatility of plastic materials has made them essential in the current era, often called the "plastic age" across most productive sectors, with global production reaching 430.9 million tons in 2024. However, their non-degradability, once considered a key advantage, is now recognized as a major environmental concern. Their extensive use since the mid-20th century combined with inadequate waste management, has led to widespread contamination of aquatic and terrestrial ecosystems. Following consumer utilization, only 9.5% of plastic is recycled, while 78% of waste ends up in landfills, soils, and waters. Conventional plastics, derived from fossil resources, comprise molecules with different chemical structures. In our project, Bioplastomics, we focus on four materials: an aliphatic polyolefin such as polyethylene (PE) of high- and low-density (HDPE and LDPE, respectively), an aromatic polymer such as polystyrene (PS), a polyester like polyethylene terephthalate (PET) and the polyolefin Polypropylene (PP). All these materials have a very high resistance to degradation, the main drivers of their decomposition in nature are abiotic factors (radiation, erosion, oxidation, etc.), whose extremely slow action leads to the release of micro and nanoplastics (100-0.001 µm) into the environment. Recent microbial research has therefore begun to investigate whether, and to what extent, microorganisms can degrade plastics in situ.

Our project aims to use cutting-edge omics techniques on Brazil environmental samples to identify new microbial strains and their active enzymes with potential for efficient plastic degradation. Samples collected from landfills, beaches, and the guts of plastic-consuming larvae were enriched on media containing the target plastic polymers.

Shotgun metagenomic analysis allowed the reconstruction of bacterial genomes belonging primarily to the phyla Pseudomonadota, Actinomycetota, and Bacteroidota — taxa previously associated with plastic polymer degradation. The proteomes encoded by these genomes are currently being compared against public databases of known plastic-degrading enzymes to identify promising enzyme candidates.

URL

https://bioplastomics.base44.app/

Functional AI-nnotation: Unlocking the “Orphan” Proteome

Poster

Damien Mornico [1], ***Natalia Pietrosemoli*** [1]

1. *Institut Pasteur*

Abstract

Genome annotation—identifying and labeling functional elements in raw DNA—underpins most downstream genomic analyses, such as variant interpretation, RNA-seq, proteomics, GWAS, and network reconstruction. Recent advances in artificial intelligence have shifted annotation from shallow statistical tools to deep-learning architectures that can grasp biological complexity. While convolutional and recurrent networks have improved structural annotation (gene identification) and graph neural networks have enriched relational annotation (metabolic pathways), functional annotation has benefitted most dramatically. Traditional pipelines infer function by transferring biological roles — often using Gene Ontology (GO) terms — from well-characterized homologs, a strategy that fails for “orphan” proteins lacking close matches. State-of-the-art protein-language models (PLMs) such as ESM-2 and ProtTrans treat amino-acid sequences as natural-language sentences, learning contextual embeddings through self-supervised pre-training on billions of residues. These embeddings encode biophysical and evolutionary information and, when combined with neuro-symbolic constraints from the GO, enable zero-shot functional prediction for proteins with no prior experimental data.

We applied PLMs to the proteomes of a non-model organism—Candidatus Arthromitus —in which ~25 % of proteins lack database matches. PLM-based annotations were benchmarked against classic homology-driven pipelines (BLAST, HMM profiles) using standard metrics and a novel validation framework suited to dark proteomes. The PLM approach reduced overall error, especially in the “twilight zone” of 20–30 % sequence identity and for orphan proteins, where it assigned GO terms unavailable to homology methods. Propagation of these predictions into RNA-seq gene-set enrichment analyses altered pathway detection, demonstrating biological relevance. Further validation employed (i) stratified benchmarking by homology depth and (ii) a functional-module-coherence test, mapping predicted GO terms onto protein-protein interaction networks and assessing their clustering coherence.

Our comparative study shows that AI-driven PLM pipelines markedly expand functional insight for understudied species, offering a robust alternative for decoding the hidden biology of the dark proteome.

Generating Chain Mappings in Large Protein Structures

Poster

Pierre Berriet [1], ***Bastien Cazaux*** [1], ***Jean-Stéphane Varré*** [1], ***Marc Lensink*** [2]

1. *Univ. Lille, CNRS, Centrale Lille, UMR 9189 CRIStAL,* ***2.*** *Univ. Lille, CNRS UMR 8576-UGSF-Unité de Glycobiologie Structurale et Fonctionnelle, 59000 Lille, France*

Abstract

Comparing protein complex conformations requires mapping chains between target and query structures. This is a computationally intensive task due to the exponential increase in possible mappings as the number of chains increases. For instance, structures comprising eight chains can generate over 40,000 potential mappings. To address this issue, we have developed a modular tool that reduces the number of mappings by taking into account sequence similarity and spatial information.

Our tool processes PDB files, identifies optimal chain mappings according to root-mean-square deviation (RMSD), and exports the results in CSV format. It uses reduction modules, namely sequence clustering, spatial clustering, pivot-based constraints and simulated annealing, to limit candidate mappings, followed by RMSD scoring. The pivot module is particularly effective for structures with rotation symmetries, while simulated annealing ensures that a solution is always found. We evaluated the tool on 5,112 predictions from five CAPRI50 structures, including the most challenging cases. With an enumeration threshold of 1,000 and a sequence similarity cut-off of 70%, the tool completed all comparisons in 29 minutes using 50 cores.

The results demonstrate that sequence and spatial clustering significantly reduce the number of mappings, enabling efficient enumeration, even for complex structures (e.g. 20-chain models). For well-modelled structures, the strong correlation between RMSD and DockQ scores confirms the tool's ability to identify optimal mappings. This approach effectively scales to large protein complexes, offering a robust solution for protein structure comparison.

URL

https://gitlab.univ-lille.fr/pierre.berriet/chain-mapping/

Genome annotations in ATLASea: using BEAURIS for their generation, FAIR handling and exploration within genomic web portals

Poster

Romane Libouban [1], ***Laura Le Goff*** [1], ***Solenne Correard*** [2], ***Mateo Boudet*** [3], ***Anthony Bretaudeau*** [4]

1. *GenOuest, Univ Rennes, INRIA, CNRS, IRISA, Rennes, France,* ***2.*** *IGEPP, INRAE, Institut Agro, University of Rennes, Rennes, France,* ***3.*** *GenOuest, Univ Rennes, Inria, CNRS, IRISA, 35000, Rennes, France,* ***4.*** *GenOuest, Univ Rennes, Inria, CNRS, IRISA, Rennes, France*

Abstract

The ATLASea project [1], led by the CNRS and the CEA, aims to sequence the genomes of 4,500 marine eukaryote species, covering the diversity of marine ecosystems in hexagonal France and overseas territories. Advances in next-generation sequencing enable unprecedented genome reconstruction accuracy, enhancing our understanding of evolutionary mechanisms, ecological interactions, and conservation strategies amid the biodiversity crisis.

Managing such vast data requires specialized tools. BEAURIS [2] is a project developed to automate, in a FAIR way, the production of functional annotation, as well as user-friendly web interfaces for visualization and comparison of genome annotations and other genomic data.

Once sequenced, the ATLASea genomes are assembled and structurally annotated by Genoscope and deposited to ENA/INSDC. BEAURIS manages, transforms, and publishes this data: each organism's data is automatically referenced in a yaml file, including lineage information and other associated metadata (e.g. TolID, Taxid, BioProject and accession identifiers) and metrics (e.g. GC content, N50 values). BEAURIS then launches the ORSON [3] functional annotation workflow whose output is loaded into a Genoboo web portal for exploration. In parallel, JBrowse2 [4] instances are created and published.

Integrated with Galaxy [5], an open-source platform that enables collaborative and reproducible workflows, BEAURIS ensures uniform processing of all genomes, whether annotated or not. The system also automates the deployment of FAIR-compliant web portals, enabling results to be shared with the scientific community.

Combining automation, standardization, and FAIR principles, BEAURIS provides a scalable and reliable framework for marine genome annotation that addresses the lack of automated, user-friendly, and FAIR-compliant tools in this field. This infrastructure has already been applied for over 300 of species and is poised to support ATLASea's goal of annotating all 4,500 target genomes. BEAURIS is currently used beyond ATLASEA (e.g. BIPAA platform [6]), and can manage other data besides genomic data.

Genome-wide DNA methylation profiles identify molecular predictors of measurable residual disease in the MIDAS Trial

Poster

Céline Chevalier [1], ***Jennifer Derrien*** [1], ***Victor Bessonneau*** [1], ***Mia Cherkaoui*** [1], ***Jill Corre*** [2], ***Aurore Perrot*** [3], ***Philippe Moreau*** [4], ***Cyrille Touzeau*** [4], ***Stéphane Minvielle*** [1], ***Florence Magrangeas*** [1], ***Eric Letouzé*** [1]

1. *Nantes Université, INSERM, CNRS, Université d'Angers, CRCI2NA,* ***2.*** *Cancer Research Center of Toulouse, INSERM, CNRS, Université Toulouse III-Paul Sabatier,* ***3.*** *Service d'Hématologie, Centre Hospitalier Universitaire de Toulouse, Institut Universitaire du Cancer de Toulouse - Oncopole,* ***4.*** *Service d'Hématologie, Centre Hospitalier Universitaire de Nantes*

Abstract

Context: DNA methylation is a critical epigenetic mechanism regulating tumorigenesis, gene expression, genomic stability, and therapeutic response (Jones et al., 2002; Berman et al., 2011). Multiple myeloma (MM), a clinically and molecularly heterogeneous hematological malignancy, is driven in part by epigenetic alterations that contribute to disease progression and drug resistance (Alzrigat et al., 2018). However, large-scale integration of DNA methylation with other omics data remains underexplored. To address this, we analyzed multiomic data from 417 newly diagnosed MM patients in the MIDAS trial (NCT04934475), combining whole-genome methylation profiles (WGEM-seq) with paired transcriptomic and clinical data, including measurable residual disease (MRD) status (Magrangeas et al., under revision).

Motivation: Our study aimed to 1.) evaluate DNA methylation as a complementary tool to refine transcriptome-based classifications (e.g., Zhan et al., 2006) and improve patient stratification, and 2.) characterize MM's epigenetic landscape at unprecedented resolution. We identified nine methylation subgroups that partially correlated with transcriptomic clusters. Strong correspondence was observed for M1 (MF), M4 (MS), and M7 (HP1/3), while M2 (CD-2) and M3 (CD-1/2) showed moderate overlap, indicating methylation captures additional heterogeneity. Notably, M8 (HP2), traditionally MRD-negative, was significantly enriched for MRD-positive samples, suggesting methylation-based stratification may redefine risk categories and improve prognostic accuracy.

Perspectives: Our findings demonstrate that whole-genome methylation profiling 1.) recapitulates known MM biology, and 2.) reveals novel epigenetic heterogeneity within transcriptomic subtypes. Future work will use MethSig (Pan et al., 2021) to identify hypermethylated promoters and Evoflux (Gabbutt et al., 2025) to track epigenetic evolution, potentially uncovering early disease biomarkers.

GenomiqueENS, the IBENS Genomics core facility

Poster

Mohammad Sufian Bin Hudari [1], Corinne Blugeon [1], Laurent Jourdren [1], Sophie Lemoine [1], Tiphaine Marvillet [1], Audrey Onfroy [1], Catherine Senamaud-Beaufort [1], Morgane Thomas-Chollier [1]

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

Abstract

The IBENS Genomics Core Facility [1-2], established in 1999, specializes in eukaryotic functional genomics, serving both classical model organisms and exotic species like jellyfish, birds, and butterflies. Our team balances wet-lab and bioinformatics expertise, supporting projects from experimental design to publication-ready data analysis. As part of the France Génomique consortium, we hold ISO 9001 certification since 2013.

Since 2016, our facility has been offering two new technologies. The first is dedicated to single-cell RNA-seq with a Chromium system from 10X Genomics based on the Drop-seq protocol. The second is focused on long read sequencing in RNA-seq. We use Oxford Nanopore Technologies sequencers, including MinION and PromethION devices to sequence full-length transcripts for isoform abundance estimation.

Over the last couple of years we have released integrations with nf-core [3], Snakemake wrappers [4] and Epi2Me [5] for ToulligQC [6], our QC tool for Oxford Nanopore sequencers. To achieve a benchmark for single-cell long read analysis tools (Keñever, Hamraoui et al., under revision [7]), we have developed AsaruSim [8] an automated workflow designed for simulating 10X Genomics single-cell long-read data. Additionally, we are currently developing a Nextflow [9] transcriptome annotation pipeline named Egzotek [10] for non-model species using RNA-seq long-reads. We also evaluate pre-trained deep learning models, such as Helixer [11], to predict gene structures directly from genome assemblies. Indeed, having a good quality annotation for exons and UTRs is mandatory with non-model species for transcriptomic applications like RNA-seq, scRNA-seq and BRB-seq.

All these developments enable us to stay at the forefront of functional genomics applications, providing our users with all the tools needed to succeed in their high-throughput experiments.

Acknowledgements

The GenomiqueENS core facility was supported by the France Génomique national infrastructure, funded as part of the "Investissements d'Avenir" program managed by the Agence Nationale de la Recherche (contract ANR-10-INBS-0009).

URL

[1] https://genomique.bio.ens.psl.eu/

[2] @genomiqueens.bsky.social

[3] https://nf-co.re/

[4] https://github.com/snakemake/snakemake-wrappers

[5] https://epi2me.nanoporetech.com/

[6] https://github.com/GenomiqueENS/toulligQC

[7] Hamroui A, Onfroy A, Senamaud-Beaufort C, Coulpier F, Lemoine S, Jourdren L, Thomas-Chollier M. A systematic benchmark of bioinformatics methods for single-cell and spatial RNA-seq Nanopore long-read data. BioRXiv 10.1101/2025.07.21.665920

[8] Hamraoui A, Jourdren L, Thomas-Chollier M. AsaruSim: a single-cell and spatial RNA-Seq Nanopore long-reads simulation workflow. Bioinformatics. 2025 Mar 4;41(3):btaf087.

[9] Di Tommaso P, Chatzou M, Floden EW, Barja PP, Palumbo E, Notredame C. Nextflow enables reproducible computational workflows. Nat Biotechnol. 2017 Apr 11;35(4):316-9

[10] https://github.com/GenomiqueENS/egzotek

[11] https://github.com/usadellab/Helixer

Geom@nnot: Environmental Metadata Enrichment from Biosample Coordinates

Poster

Mélodie Fleury [1], ***Elea Pauliat*** [2], ***Paul Tissot*** [2], ***Luca Nesterenko*** [3], ***Stephane Delmotte*** [4], ***Maël Rimeur*** [1], ***Romain Delunel*** [5], ***Julien DELLINGER*** [2], ***Caroline Leroux*** [6], ***Jérome Lejot*** [5], ***Romuald Marin*** [7], ***Matis Zouari*** [8], ***Christophe Blanchet*** [8], ***Dominique Guyot*** [2], ***Christine Oger*** [2], ***Damien de Vienne*** [9], ***François Mialhe*** [5], ***Hussein Anani*** [10], ***Laurence Josset*** [10], ***Jocelyn Turpin*** [6], ***Vincent Navratil*** [2], ***Oldrich Navratil*** [5]

1. *PRABI, Rhône-Alpes Bioinformatics Center, Université Lyon 1, 43 bd du 11 novembre 1918, Villeurbanne CEDEX 69622, France.*, ***2.*** *PRABI, Rhône-Alpes Bioinformatics Center, Université Lyon 1, 43 bd du 11 novembre 1918, Villeurbanne CEDEX 69622, France,* ***3.*** *Université Claude Bernard Lyon 1,* ***4.*** *Universite Claude Bernard Lyon 1, LBBE, UMR 5558, CNRS, VetAgro Sup, Villeurbanne 69622, France,* ***5.*** *CNRS 5600 EVS, Université Lumière Lyon 2,* ***6.*** *IVPC UMR754, INRAE, Universite Claude Bernard Lyon 1, EPHE, PSL Research University, 69007, Lyon, France,* ***7.*** *CNRS, UAR 3601 ; Institut Français de Bioinformatique, IFB-core, 7 rue Guy-Môquet, F-94800 Villejuif, France,* ***8.*** *IFB-core CNRS UAR3601,* ***9.*** *Laboratoire de Biométrie et Biologie Évolutive, Lyon, France,* ***10.*** *Hospices civils de Lyon*

Abstract

Background: Linking nucleotide sequences (or samples) in public repositories with environmental metadata (*e.g.*, climate, biome, population density, GDP) are essential from a One Health perspective. However, these metadata are often incomplete, inaccurate, poorly standardized, and prone to human error during submission. Existing solutions mainly rely on manual curation from publications and provide metadata with limited diversity and completeness. The growing number of global geospatial databases (*e.g.*, NaturalEarth, HydroSHEDS) now easily accessible (*e.g.*, through Google Earth Engine), offers new opportunities to fill data gaps and enrich metagenomic datasets with geoenvironmental variables. Given the diverse origin of these sources (e.g., field, satellite sensors, modelling) and their heterogeneous spatial resolutions and completeness, the Virome@tlas project, through a collaboration between bioinformaticians and geomaticians, has developed a rigorous framework to automate the recontextualisation of metagenomic sequencing data

Results: Virome@tlas harmonized dataset includes 38M geolocated viral GenBank nucleotide sequences and/or biosamples from INSDC and GSA. A total of 750,000 unique coordinates were processed by Geom@nnot, a geomatics Python tool, achieving a 50-fold compression rate and improving annotation scalability (Fig. 1A). Each coordinate is automatically assigned to a spatial scale (*i.e.*, from plot scale to country). Georeferenced nucleotide sequences (or samples) were next recontextualized using environmental datasets (*e.g.*, atmosphere, hydrosphere, pedosphere, biosphere and anthroposphere). Geospatial datasets were selected based on traceability, resolution, consistency with location accuracy, documentation availability, and scientific validation (Fig. 1B). Finally, enriched metadata were computed (Fig 1C) and harmonized using state-of-the-art ontologies and controlled vocabularies from bioinformatics and geomatics (*e.g.*, ENVO, GOLD).

Conclusion: We developed a Python tool able to efficiently inform environmental metadata for millions of nucleotide sequences (or samples) using diverse geospatial databases. These new metadata are integrated into the Virome@tlas datalake and visualization platform following FAIR principles, opening new avenues to explore virus-host-environment relationships (https://viromeatlas.univ-lyon1.fr/).

URL

https://viromeatlas.univ-lyon1.fr/

GPU pipeline and interactive interface for large-scale single-cell data analysis and visualisation

Poster

Astrid Delépine [1], ***Lilia Younsi*** [1], ***Yoann Martin*** [1], ***Benjamin Saintpierre*** [1]

1. *Plateforme Bioinformat'ic Institut Cochin, Paris, France Institut Cochin, Université Paris Cité, CNRS-8104, Inserm U1016, Paris, France.*

Abstract

The growing prevalence of single-cell RNA-seq projects has introduced significant challenges in data processing and time management. The increasing complexity of the data and the size of the datasets sometimes make traditional scRNA-seq analysis pipelines slow or inefficient. These pipelines, often limited by central processing unit (CPU) resources, are no longer sufficient to handle large-scale experiments effectively. Additionally, the interpretation and customization of results remain accessible only to experts.

To answer these challenges, we implemented a parallel computing solution with Graphics Processing Unit (GPU), integrating Compute Unified Device Architecture (CUDA) and using optimised library as Scanpy and RAPIDS-singlecell (Scverse) to aggregate up to one million cells and significantly accelerate data processing. This solution makes the whole pipeline execution (from running quality control to UMAP annotation) up to 10 times faster, reducing analysis times without compromising results quality. The execution time was subsequently reduced from one hour to six minutes for samples comprising 500,000 cells. In addition, we began developing two web interfaces using the Python framework Panel. One interface will allow bioinformaticians to directly assess the impact of parameter changes on their pipeline, while the other will provide an interactive visualization interface enabling researchers to independently explore data that has already been analysed by the facility.

Our GPU-accelerated pipeline and user-friendly interfaces will provide a robust solution to the challenges of scalability and accessibility in single-cell RNA-seq analysis. By decreasing execution times and facilitating researchers' interaction with their data, this work sets a foundation for more efficient and collaborative exploration and data sharing.

HaploExplore, a software specifically designed for the detection of minor allele (MiA-) haploblocks

Poster

Samuel HIET [1], ***Matilde Manetti*** [1], ***Myriam Rahmouni*** [1], ***Jean-Louis Spadoni*** [1], ***Alice Dobiecki*** [1], ***Marco Lamanda*** [1], ***Maxime Tison*** [1], ***Taoufik Labib*** [1], ***Cristina Giuliani*** [2], ***Sigrid Le Clerc*** [1], ***Jean-François Deleuze*** [3], ***Jean-François Zagury*** [1]

1. *Laboratoire Génomique, Bioinformatique, et Chimie Moléculaire, EA7528, Conservatoire National des Arts et Métiers,* ***2.*** *Laboratory of Molecular Anthropology, Department of Biological, Geological and Environmental Sciences, University of Bologna,* ***3.*** *Laboratory for Genomics, Foundation Jean Dausset - CEPH*

Abstract

Haplotype blocks in the genome are informative of evolutionary processes and they play a pivotal role in describing the genomic variability across human populations and susceptibility/resistance to diseases. Several software have been developed for haplotype block detection, but they do not distinguish between the impacts of major and minor single nucleotide polymorphism (SNP) alleles. In this study, we present a powerful haploblock detection software, specifically designed for identifying haploblocks associated with SNP minor allele haploblocks (MiA-haploblocks). These haploblocks are particularly important as they can significantly influence phenotypic traits, offering a novel approach for studying genetic associations and complex traits.

HaploExplore operates on VCF files containing phased data, exhibiting rapid processing times, and generating user-friendly outputs. Results converge when analyzing populations of 100 individuals or more. A comparative analysis of HaploExplore against other haploblock detection software revealed its superiority in terms of either simplicity, flexibility, or speed, with the unique capability to target minor alleles. HaploExplore will be very useful for evolutionary genomics and for GWAS analysis in human diseases, given that the effects of genetic associations may accumulate within a specific haploblock.

URL

1. https://doi.org/10.1093/nargab/lqaf186

2. https://doi.org/10.6084/m9.figshare.30518498

How Data Pre-processing Shapes Conclusions in Metagenomics: A Reproducible Benchmark to Guide Microbiome Analysis

Poster

Emile Mardoc [1], ***Maxence Klock*** [1], ***Xavier Raffoux*** [1], ***Julie Aubert*** [2], ***Christelle Hennequet-Antier*** [3], ***Mathilde Sola*** [1], ***Emmanuelle Le Chatelier*** [1], ***Nicolas Maziers*** [1], ***Florence Thirion*** [1], ***Florian Plaza Oñate*** [1], ***Giacomo Vitali*** [1], ***Lindsay Goulet*** [1], ***Mahendra Mariadassou*** [3], ***Mathieu Almeida*** [1], ***Magali Berland*** [1]

1. *INRAE MGP,* ***2.*** *INRAE MIA,* ***3.*** *INRAE MaIAGE*

Abstract

The human gut microbiome, contributing to digestion, immunity, and the synthesis of bioactive metabolites, is investigated through shotgun metagenomics to uncover its associations with human diseases. However, metagenomic data analysis requires a cascade of preprocessing steps that aim to correct technical biases (sequencing depth, paired-end reads) and address the statistical properties of the data (compositional, non-Gaussian, with high variability). Choices such as the rarefaction threshold, the normalization method, or the transformation of abundance data constitute “researcher degrees of freedom” which can markedly affect downstream biostatistical results and interpretations.

We present a fully reproducible R-based benchmark to evaluate the relevance and impact of preprocessing using six large (n > 300) public datasets spanning distinct clinical and geographic settings. Pre-processing effects are evaluated for each downstream analysis type: (i) alpha-diversity estimation, (ii) multivariate ordination (PCoA, NMDS), (iii) differential abundance testing, (iv) supervised classification (random forests), and (v) inference of microbial interaction networks. The impact is evaluated in terms of robustness of the results, assessed by the stability of statistical estimates across multiple subsampling iterations.

Unlike previous benchmarks, our framework focuses on the specific characteristics of shotgun metagenomic data and extends beyond differential abundance to cover a broad spectrum of analytical goals. The results yield practical recommendations for selecting the most appropriate pre-processing x analysis combinations according to biological context and analytical objective. This work embodies an open and reproducible science, representing a step toward standardized microbiome data analysis.

How metagenomic analysis strategy shapes functional inference? Metabolic landscapes from Le French Gut Cohort

Poster

Toubal Sarah [1], ***Le French Gut Consortium*** [2], ***Magali Berland*** [3], ***Clémence Frioux*** [1], ***Florian Plaza Oñate*** [3]

1. *Inria, Université de Bordeaux, INRAE,* ***2.*** *https://lefrenchgut.fr/le-projet/les-partenaires/,* ***3.*** *INRAE MGP*

Abstract

Background The human gut microbiome plays a key role in host health by contributing to digestion, immune regulation, and metabolic processes. Alterations in microbial community composition have been linked to numerous diseases, motivating large-scale metagenomic studies to better characterize microbial diversity and function. Projects such as *Le French Gut* aim to describe gut microbiota structure at the population level (currently n = 10,000 individuals) using shotgun metagenomic sequencing. However, different analytical strategies may strongly influence genome reconstruction, functional annotation, and downstream analyses.

Results We analyzed 105 fecal metagenomes from the *Le French Gut* cohort using both a reference-based profiling strategy and a de novo assembly-based protocol to retrieve metagenome-assembled genomes (MAGs). For the de novo approach, we additionally evaluated the impact of sequence down-sampling on MAG reconstruction. Each sample was characterized as a collection of reference genomes, de novo MAGs, and as a gene catalog. For each strategy, metabolic pathways and metabolic networks were reconstructed using several existing tools, revealing notable differences in inferred functional potential depending on the chosen metagenomic strategy.

Conclusions Our results demonstrate that the choice of metagenomic analysis strategy significantly impacts functional inference. De novo and reference-based approaches provide complementary views of the gut microbiota. We further analyzed reconstructed metabolic profiles in the context of the extensive health, lifestyle, and dietary metadata associated with the dataset. Overall, our findings highlight the importance of methodological choices in metagenomic studies and provide a framework for integrating microbial genomic data with clinical, dietary, and lifestyle information in large population cohorts.

ICEs and IMEs Delineation : Leveraging Pangenome and Machine Learning Approaches

Poster

Mamadou Aliou Diallo [1], ***Thomas Lacroix*** [1], ***Hélène CHIAPELLO*** [1], ***Guillaume GAUTREAU*** [1]

1. *INRAE*

Abstract

Integrative Conjugative Elements (ICEs) and Integrative Mobilizable Elements (IMEs) are major contributors to horizontal gene transfer in bacteria. These mobile genetic elements play crucial roles in disseminating antibiotic resistance genes, virulence factors, and other adaptive traits. However, detecting, delineating, and annotating these elements within chromosomes remains complex due to several intrinsic challenges.

One difficulty lies in their modular structure and diversity in size and gene content. They can aggregate into intricate structures, such as accreted or matryoshka-like configurations, where multiple ICEs or IMEs are nested within one another. Additionally, their precise delineation is hampered by the unreliable detection of the small and often degenerate direct repeats left by their integration.

Current state-of-the-art strategies for delineating ICEs and IMEs often rely on the availability of user-provided closely related genomes, some carrying the mobile element while others do not. By comparing synteny across these genomes, these methods then infer element locations. While effective, these approaches are labor-intensive and not always feasible.

This work presents preliminary results on two complementary approaches toward automating the detection, delineation, and annotation of ICEs and IMEs in a generic framework. The first relies on an explicit approach combining pangenome analysis and protein signatures. We use both ICEscreen for its ability to detect signature proteins, and PPanGGOLiN for its ability to classify genes in persistent and non-persistent partitions. The second methodology uses transfer learning based on the Evo 1.5 large genomic model to learn genomic signatures of ICEs and IMEs from datasets of annotated genomes. Comparison of these approaches suggests strong complementarity. The explicit method provides accurate delineation and interpretability, whereas the Evo1.5-based model shows promising generalization across taxa but with lower interpretability and specific computational resources (recent GPUs). Future work will focus on integrating the strengths of both strategies into a dedicated software solution.

Identification of microorganisms in dairy systems using shotgun metagenomic data

Poster

Oriane Lamy [1], ***Solène Pety*** [1], ***Fiona Bottin*** [2], ***Sébastien Theil*** [2], ***Pierre Nicolas*** [1], ***Guillaume Kon Kam King*** [1], ***Céline Delbès*** [2], ***Anne-Laure Abraham*** [1]

1. *Université Paris-Saclay, INRAE, MaIAGE, 78350, Jouy-en-Josas, France,* ***2.*** *UMR 545 Fromage, Université Clermont Auvergne, INRAE, VetAgro Sup, Aurillac, France*

Abstract

In the context of climate change, agroecological methods are key for developing more sustainable and resilient food systems, especially for the production of raw milk cheese, whose quality depends heavily on environmental microbial diversity. The TANDEM (Transfers iN Dairy systEM) project addresses this issue by studying how microbial ecosystems respond to changes in cow feed and how microorganisms are transmitted throughout the dairy food chain. A total of 199 samples were collected from seven ecosystems (soil, grass, rumen, feces, milk, cheese rind, and cheese core) and analysed using shotgun metagenomic methods. The obtained data can be used to study community composition and identify shared microorganisms at the lineage level.

To this end, a catalog of reference genomes representative of the project's ecosystems, must first be built. However, the data of the project allowed to assemble only 266 MAGs (Metagenomics Assembled genomes), due to high species diversity in some ecosystems, which is insufficient. Furthermore, there are no reference catalogs in public databases for all these ecosystems, so a dedicated catalog was built using a generalist database (RefSeq) and genomes collected from several past experiments (on cow rumen and feces, cheese, and human gut). Two such catalogs were created, and compared after reads alignment using BWA-MEM (Burrows–Wheeler Aligner with Maximal Exact Matches). The second one was selected for downstream analyses as it yielded better results. The alignment data will then be analysed using a dedicated Snakemake workflow called INTERSTICE. It takes into account shared single nucleotide polymorphisms (SNPs) between samples to compute similarity indices, based on Nei's distance, adapted for metagenomic samples. These distances can help characterize bacterial flows across the dairy production chain by studying intraspecies diversity at two levels: the presence of multiple strains within the same environment, and whether these strains are shared across environments.

IFB-Biosphère Cloud, Multi-Cloud Infrastructure for Life Sciences

Poster

Christophe Blanchet [1], ***Mateo Boudet*** [2], ***Guillaume Brysbaert*** [3], ***Micael Calvas*** [4], ***Stephane Delmotte*** [5], ***Hervé Gilquin*** [4], ***Nadia Goué*** [6], ***Jean François Guillaume*** [7], ***Antoine Mahul*** [8], ***Jérôme Pansanel*** [9], ***Bruno Spataro*** [5], ***Cyrille Toulet*** [10], ***Matis Zouari*** [1]

1. *IFB-core, Institut Français de Bioinformatique (IFB), CNRS, INSERM, INRAE, CEA, 94800 Villejuif, France.,* ***2.*** *GenOuest, Univ Rennes, Inria, CNRS, IRISA, 35000, Rennes, France,* ***3.*** *Univ. Lille, CNRS UMR 8576-UGSF-Unité de Glycobiologie Structurale et Fonctionnelle, 59000 Lille, France,* ***4.*** *CBPsmn, Centre Blaise Pascal de simulation et modélisation numérique, Ecole Normale Supérieure de Lyon, France,* ***5.*** *Universite Claude Bernard Lyon 1, LBBE, UMR 5558, CNRS, VetAgro Sup, Villeurbanne 69622, France,* ***6.*** *Plateforme AuBi, Mésocentre Clermont-Auvergne, Université de Clermont-Ferrand, Aubière, France,* ***7.*** *GLiCID, BiRD, UMS BioCore (Inserm US16 et UAR CNRS 3556), UFR Médecine et Techniques Médicales, Nantes Université, CHU Nantes, France,* ***8.*** *Mésocentre Clermont-Auvergne, Université de Clermont-Ferrand, Aubière, France,* ***9.*** *Université de Strasbourg, CNRS, IPHC UMR 7178, F-67000 Strasbourg, France,* ***10.*** *Mésocentre régional, Université de Lille, France*

Abstract

The **IFB-Biosphere cloud**, provided by the French Institute of Bioinformatics (IFB), offers scientists a wide range of digital services for Biology applications. It enables them to deploy on-demand digital research environments in the form of **virtual machines (VMs)** and **containers** with integrated bioinformatics tools for the analysis of biological data. The **Biosphere infrastructure** provides cloud services at various levels: for **scientific users**, **bioinformatics application developers**, **cloud administrators**, and **Life Sciences trainers**.

The **Biosphere infrastructure** is a national federation of **8 clouds** located within IFB bioinformatics platforms, some of which are developed in collaboration with national mesocenters or data centers. The software configuration is based on open standards such as **OpenStack** for virtual machines, **Docker** for containers, **CEPH** and **ZFS** coupled with Manila for shared storage between VMs. The infrastructure includes over **10,000 vCPUs** and various GPU models, offering a variety of VM templates—ranging from a few vCPUs and a few GB of memory to **BigMem VMs** (up to 4TB of RAM) and VMs with many vCPUs (up to 255 vCPUs per VM).
The **Biosphere multi-cloud web portal enables users to deploy scientific environments across IFB clouds through a unified interface**. This Biosphere portal offers several core services:

- The **RAINBio catalog** of reproducible bioinformatics environments,
- A **dashboard** for managing VMs and data (public and project-specific),
- A **user authentication service** based on the Renater identity federation.

The bioinformatics software environments available in the **RAINBio catalog** include widely used scientific tools such as **R/RStudio/Shiny, Jupyter, Nextflow, and Snakemake**, as well as other thematic tools developed by domain experts.

The **Biosphere cloud** acts as a national digital infrastructure dedicated to biological research and scientific training programs, such as thematic schools, university courses, technical workshops, and hackathons. It supports these initiatives by allocating dedicated CPU/RAM and storage resources, as well as specific services.

URL

https://biosphere.france-bioinformatique.fr

Improving accessibility of machine learning models in bioinformatics

Poster

Pauline Le Corre [1], ***Anthony Bretaudeau*** [2], ***Yann Le Cunff*** [1]

1. *GenOuest, Univ Rennes, INRIA, CNRS, IRISA, Rennes, France,* ***2.*** *GenOuest, Univ Rennes, Inria, CNRS, IRISA, Rennes, France*

Abstract

Machine learning (ML) is increasingly used in bioinformatics for various applications. While many models are available on repositories like HuggingFace, their use is not always easy and well documented. To ensure reproducibility and accessibility, it is essential to provide user-friendly tools for running models, allowing a larger community, including researchers without advanced computational skills, to benefit from them.

We explore various strategies to improve the usability of ML models, using EnzBERT as a case study. EnzBERT is a transformer-based model for predicting enzyme functional annotation from the protein sequences. Initially, running the model required downloading the model and manually executing the code. To address this, we implemented 3 approaches. The first approach is to create a Bioconda package which packages the code needed to run the model and enables an easier execution. Using this package, we also developed a galaxy tool for running the model. Galaxy is an open-source platform providing bioinformatics tools via a web interface. Finally, the last explored approach was to create a web-based user interface and an API running the model. We will present the benefits of each approach based on the ease of use for end-users and on the ease of implementation for developers.

Additionally, we emphasize the importance of transparency, reproducibility, and adherence to FAIR principles for ML models. Several recommendations exist, including the DOME guidelines, tailored for ML in biology. DOME is an acronym for Data, Optimization, Model, Evaluation, corresponding to the scopes covered by the guidelines. DOME provides a framework of key questions to address during the model development ensuring best practices.

Several strategies exist for sharing ML models in bioinformatics, each requiring a balance between user accessibility and developer practicality. By adopting frameworks such as the DOME guidelines, models can be made more transparent, reproducible, and aligned with FAIR principles.

Improving viral protein clustering using both diversified protein profiles and structural information

Poster

Quentin Nugier [1], ***George Bouras*** [2], ***Clovis Galiez*** [3], ***Marie-Agnès Petit*** [1], ***Francois Enault*** [4]

1. *INRAE, Micalis department,* ***2.*** *University of Adelaide, medical school,* ***3.*** *Université Grenoble Alpes,* ***4.*** *Université Clermont-Auvergne, LMGE*

Abstract

Viruses are abundant, ancient, and rapidly evolving biological entities. As a result, viral proteins are highly diverse, making the identification of homologous relationships both essential for phylogenetic inference and functional annotation and particularly challenging. Viral genome annotation largely relies on transferring annotations from large curated protein family databases such as PHROG, making improvements to these resources critical.

The most sensitive sequence analysis methods are based on probabilistic models (HMMs) but strongly depends on the sequence diversity used to build them. Here, proteins from reference viral genomes were first clustered using classical sequence-based approaches. The resulting profiles were then enriched with tens of millions of viral metagenomic sequences, substantially increasing sequence diversity across most families. These enriched models detected more than three times as many homologous relationships as models built from the original datasets.

Protein families were subsequently clustered based on all detected links and further unified using structural prediction and comparison. Using this combined strategy, 1.42 million proteins grouped into only 56,560 families, compared to 200,018 and 135,048 families obtained using traditional sequence-based and raw model comparison methods.

The enriched sequence approach was particularly effective at revealing evolutionary links between small proteins, whereas structural comparisons were more effective for highly structured proteins such as head and tail proteins. Together, these complementary approaches reveal deep evolutionary relationships and provide a more accurate view of viral protein diversity and viral evolutionary history.

The resulting protein families will be annotated and made accessible through an online database.

In Silico Prediction of Transcription Factor Binding Sites in Proximal Promoter Regions Using TSS-Relative Positional Enrichment

Poster

Margot CORREA [1]***, Guichard Cécile*** [2]***, Guillem Rigaill*** [3]***, Véronique Brunaud*** [2]

1. *Université Paris-Saclay, CNRS, Univ Evry, Laboratoire de Mathématiques et Modélisation d'Evry, 91037, Evry, France.,* ***2.*** *1-Université Paris-Saclay, CNRS, INRAE, Univ. Evry, Institute of Plant sciences Paris-Saclay (IPS2), 91190, Gif sur Yvette; 2-Université de Paris Cité, Institute of Plant sciences Paris-Saclay (IPS2), 91190, Gif-sur-Yvette, France,* ***3.*** *1-,2-,3-Université Paris-Saclay, CNRS, Univ Evry, Laboratoire de Mathématiques et Modélisation d'Evry, 91037, Evry, France; 4-Université Paris-Saclay, AgroParisTech, INRAE, UMR MIA Paris-Saclay, 91120, Palaiseau, France.*

Abstract

Background

Transcription factor binding sites (TFBS) are key regulatory elements controlling gene expression and are often enriched in proximal promoter regions relative to the transcription start site (TSS). However, identifying biologically functional TFBS and understanding their interaction with transcription factors remain major challenges. This is despite massive progress due to AI models for bioinformatic and because experimental methods cover only a fraction of TF. We hypothesized that positional enrichment relative to the TSS constitutes an informative constraint for TFBS prediction.

Results

Using an *in silico* approach, we searched for motifs enriched at specific distances from the TSS within the proximal promoters of *Arabidopsis thaliana* stress response genes. This strategy led to the identification of preferentially located motifs (PLM). Among them, 91% correspond to TFBS previously annotated in the JASPAR database, demonstrating that the PLM-motifs are indeed TFBS. In addition, 99.6% of newly predicted motifs showed a significant association with TF binding data, suggesting that they represent biologically relevant candidate TFBS not yet annotated.

Conclusion

These results highlight positional enrichment near the TSS as a strong indicator of functional TFBS and support the integration of TSS-distance constraints to improve motif discovery. In particular for poorly studied species where binding data is scarce.

Inference of ligand–receptor interactions guiding neuronal wiring in the developing mouse somatosensory cortex

Poster

Antoine De Chevigny [1], ***Tangra Draia-Nicolau*** [1], ***Rémi Mathieu*** [1], ***Léa Corbières*** [1], ***Annousha Govindan*** [1], ***Bensa Vianney*** [1], ***Emilie Pallesi-Pocachard*** [1], ***Lucas Silvagnoli*** [1], ***Alfonso Represa*** [1], ***Carlos Cardoso*** [1], ***Ludovic Telley*** [2]

1. *INMED,* ***2.*** *Université Claude Bernard Lyon 1*

Abstract

The assembly of cortical circuits relies on tightly regulated molecular interactions between developing neuronal populations, yet the ligand–receptor (LR) logic governing these processes remains incompletely understood. To systematically infer LR-mediated cell–cell communication during cortical development, we integrated single-cell and single-nucleus RNA sequencing data from the mouse somatosensory cortex across 17 developmental stages spanning embryonic to adult life. We generated a comprehensive transcriptomic dataset of cortical neurons by profiling key postnatal windows of circuit formation, including P0–P2 (radial migration and laminar allocation), P5–P8 (programmed cell death of glutamatergic and GABAergic neurons), and P16–P30 (synaptic refinement and circuit maturation). These data were combined with previously published embryonic (E11.5–E18.5) and adult datasets, including ganglionic eminence–derived populations, to capture early transcriptional programs of future cortical interneurons.

By cross-referencing cell-type-resolved expression profiles with our curated ligand–receptor database, we inferred dynamic and stage-specific interaction networks across excitatory and inhibitory neuronal subtypes. This analysis revealed temporally restricted ligand–receptor pairs associated with neuronal migration, subtype specification, and synaptic wiring, highlighting context-dependent molecular programs underlying cortical circuit assembly. To facilitate exploration of these results, we developed an interactive Shiny application that enables users to query ligand–receptor interactions across developmental stages and neuronal populations.

Currently, my thesis project builds directly on this work by incorporating MERFISH spatial transcriptomic data, adding a clear spatial dimension to the analysis. This extension focuses on mouse brain timepoints at P0, P5, P8, and P30.

This work provides a comprehensive resource for investigating molecular communication underlying somatosensory cortical wiring and offers a publicly accessible resource for hypothesis generation.

URL

https://www.nature.com/articles/s41467-025-68059-8

Inferring Cell Fate Trajectories in Time-Resolved Metabolic RNA Labeling data

Poster

Anna Audit [1], ***Gabriel Peyré*** [2], ***Laura Cantini*** [3]

1. *Machine Learning for Integrative Genomics Group,Institut Pasteur, CNRS UMR 3738,* ***2.*** *DMA de l'Ecole Normale Supérieure, CNRS, Ecole Normale Supérieure, Université PSL,* ***3.*** *Institut Pasteur*

Abstract

Single-cell RNA sequencing technology has revolutionized the biomedical field by enabling high-resolution analysis of the transcriptional content of individual cells. However it offers only a static snapshot, missing transcriptional dynamics. Metabolic RNA labeling is a new sequencing technique that allows one to distinguish total and newly synthesized RNA [1].

Labeling newly synthesized RNA reveals the direction of individual cell evolution, which traditional sequencing cannot capture. When combined with temporal sequencing, it provides insights into both population-level changes and each cell's developmental trajectory. This is particularly valuable for trajectory inference, which aims to recover the underlying continuous dynamics from static snapshots of cell distributions. Current methods do not leverage this additional information.

We propose a method for single-cell trajectory inference that leverages information from time-resolved RNA labeling. We hypothesize that cells evolve to minimize an underlying potential function and aim to learn this potential landscape, modeled by a simple neural network [2], [3]. Our method takes as input cell populations profiled at multiple time points, together with their total and labeled RNA measurements, and predicts the evolution of the cell distribution. We use optimal transport to evaluate prediction accuracy and align the gradient of the potential with prior knowledge derived from the labeled RNA. After training, the learned energy landscape identifies key genes and transcription factors and predicts a cell's future evolution. We benchmark our method on its ability to predict unseen transcriptional states and recover coherent cell-type transitions. We also investigate deeper the dynamics of HTC116 colorectal cancer cells being treated with a DNA demethylating agent as well as neuron differentiation of mouse embryonic stem cells .

InSillyClo: How to make large-scale golden gate cloning and MoClo workflows user-friendly and reproducible

Poster

Henri Galez [1], ***Bryan Brancotte*** [2], ***Juliette Bonche*** [2], ***Julien Fumey*** [2], ***Sara Napolitano*** [1], ***Gregory Batt*** [1]

1. *Institut Pasteur, Inria, Université Paris Cité, 75015 Paris, France,* ***2.*** *Institut Pasteur, Université Paris Cité, Bioinformatics and Biostatistics Hub, F-75015 Paris, France*

Abstract

Large-scale cloning campaigns are essential in systems and synthetic biology, where constructing numerous genetic circuit variants demands efficient molecular biology workflows. Golden Gate assembly and Modular Cloning (MoClo) have emerged as powerful strategies to streamline these efforts. However, they involve repetitive and error-prone tasks such as plasmid design documentation, primer validation, dilution calculations, and gel band prediction—particularly burdensome at scale. To address this challenge, we developed InSillyClo, an open-source tool available both as a web application and a command-line interface (CLI) for comprehensive cloning campaign planning and simulation.

The CLI enables automation of cloning campaigns through assembly simulation, dilution calculation, and virtual gel prediction. Users can specify input parts, define assembly parameters, and simulate agarose gels after PCR and restriction digestion. All steps are executed programmatically, supporting integration into automated pipelines and ensuring experimental reproducibility by eliminating manual errors and enforcing standardized procedures across large datasets. The CLI is a pure-python package, named insillyclo.

While the CLI targets advanced users seeking automation, the companion web interface prioritizes accessibility and ease-of-use. Built using UX methodology, it guides users through structured workflows with minimal learning curve. Interactive forms, visual feedback, and contextual help make cloning design approachable for students and researchers alike. Together, the CLI and web interface offer flexible use of the same core engine, ensuring reproducible results both computationally and experimentally.

InSillyClo addresses a critical need for scalable, reliable cloning tools by combining user-centered design and robust automation. By facilitating accurate planning and simulation of complex cloning campaigns, it reduces repetitive task burden while improving experimental reliability. The integration of MoClo's typing system enhances design clarity and prevents errors, while automated generation of dilution schemes and validation gels streamlines both workflow design and wet-lab execution. Available at https://insillyclo.pasteur.cloud/, InSillyClo provides the community an accessible, open-source solution for modern cloning challenges.

URL

https://pubs.acs.org/doi/10.1021/acssynbio.5c00553?fig=tgr1&ref=pdf
https://insillyclo.pasteur.cloud/
https://gitlab.pasteur.fr/insillyclo/insillyclo-cli

Integrative gene network analysis of genome-wide association data in myalgic encephalomyelitis / chronic fatigue syndrome

Poster

Katia Antonenko [1], ***Giann Karlo Aguirre-Samboni*** [1], ***Florian Massip*** [1], ***Chloé-Agathe Azencott*** [1]

1. *Mines Paris PSL*

Abstract

Myalgic encephalomyelitis / chronic fatigue syndrome (ME/CFS) is a common though poorly understood disease affecting millions of people worldwide. The biological mechanisms underlying ME/CFS remain largely unclear, no effective treatments currently exist, and the disease disproportionately affects females and is frequently triggered by acute infection. However, no satisfactory mechanistic explanation for either factor has been established.

In the DecodeME study, the first genome-wide association studies (GWAS) were performed on a large cohort of cases (15,579) and controls (259,909) with European genetic ancestry. Eight loci were reported to be significantly associated with ME/CFS, three of which are proximate to genes involved in the response to viral or bacterial infection, consistent with the known infection trigger. The initial findings also suggest that both immunological and neurological processes contribute to the genetic risk of ME/CFS.

However, GWAS is by design limited to individual SNP-phenotype associations, and gene-gene interactions are largely overlooked by this framework. Polygenic diseases such as ME/CFS are likely shaped by the coordinated activity of multiple genes within shared biological pathways, rather than by isolated variants alone. Gene network methods integrating pathway and interaction data have therefore been developed to refine and enrich classical GWAS signals, and combining multiple such methods has been shown to improve statistical power and interpretability, with successful applications in breast cancer and psoriasis.

In the present study, we re-analyse the DecodeME GWAS summary statistics and apply a combination of gene network methods across curated pathway databases and experimentally derived protein interaction networks. This integrative approach yields a robust consensus of genes potentially involved in ME/CFS pathogenesis. The network analysis partly recovers the genes from the original study and additionally identifies multiple previously unreported pathways and candidate genes related to immune regulation and neurological function.

Inter-individual variability in transcriptomes: what methods can already be used and why should it be analysed?

Poster

Simon Thiry [1], ***Fabrice Teletchea*** [2], ***Elise Billoir*** [1], ***Sophie Prud'homme*** [1]

1. *Université de Lorraine, CNRS, LIEC, F-57000 Metz, France,* ***2.*** *Université de Lorraine, CNRS, LIEC, F-54000 Nancy, France*

Abstract

While omics datasets do contain a vast amount of information, making the most of it can prove complex. Majority of studies using transcriptomic data are limited to differential expression (DE) analysis, i.e. identifying differences in the central tendency of transcript / gene abundance between two conditions, or sometimes along a gradient of conditions. On the other hand, inter-individual variability of transcript abundance receive much less attention from researchers, and when it does, it frequently is via measures of statistical dispersion (e.g. standard deviation, median absolute deviation, coefficient of variation) that are unsuited for bulk RNA-Seq data and its characteristics (i.e. mean-variance dependency, skewed, overdispersed). Especially in fields where the use of omics-based approaches is relatively less established, like ecophysiology or ecotoxicology, there is a need to understand what insights may be gained by investigating gene transcription beyond differences in average abundances, and to review methodological approaches available for studying inter-individual variability. Although there are no standard methods as for DE analysis, which is a requirement for widespread adoption of inter-individual variability analysis, a few tools for differential variability (DV) and differential distribution (DD) detection have been developed during the last decade. This poster aims to: i. describe how studying inter-individual variability could enhance our understanding of plasticity mechanisms and populations acclimation to changing environments; ii. list the challenges of analysing transcriptomic variability posed by RNA-Seq data properties; iii. present existing methods & tools for DV and DD analysis.

International Society for Computational Biology Student Council Regional Student Group France (RSG France) : Association of Young Bioinformaticians of France (JeBiF)

Poster

Elisabeth Hellec [1], ***Benjamin Loire*** [2], ***Jérémy Rousseau*** [3], ***Yanis Asloudj*** [4], ***Magis Papail*** [4], ***Alexandre Lerévérend*** [4], ***Walid Sabeur*** [4], ***Célia Brahimi*** [5], ***Vinh-Son Pho*** [6]

1. *ANSES,* ***2.*** *Neurology Therapeutic Area, R&D Servier Paris-Saclay Institut,* ***3.*** *Museum Nation,* ***4.*** *Association des Jeunes Bioinformaticien-ne-s de France (RSG France - JeBiF),* ***5.*** *Aix Marseille University, INSERM, MMG UMR 1251,* ***6.*** *Sorbonne Université, CNRS, IBPS, Department of Computational, Quantitative and Synthetic Biology (CQSB), UMR7238*

Abstract

JeBiF, or “Jeunes Bioinformaticien·ne·s de France,” (young bioinformaticians of France) is a non-profit organization founded in 2008, representing the French branch of the International Society for Computational Biology Student Council (ISCB-SC), which oversees 26 local branches worldwide. The primary objective of JeBiF is to support the development of the next generation of bioinformaticians by providing networking opportunities, career guidance, and professional development. The organization also promotes computational biology and bioinformatics through science outreach initiatives.

- BioStream: BioStream is a live-streaming series dedicated to bioinformatics, featuring discussions, researcher interviews, and career insights. Ten episodes are currently available in replay on our YouTube channel, covering topics such as bioinformatics networks, forensic science, marine bioinformatics, and bioinformatics platforms. A podcast version will soon be available to broaden access to these discussions.
- JeBiF@JOBIM: JeBiF@JOBIM is an annual workshop organized before JOBIM, providing a space for the community to gather for a day to discuss and collaborate on a common project. The latest edition focused on science communication in bioinformatics, featuring a roundtable discussion, a short training session on science outreach, and an afternoon dedicated to designing four interactive mediation workshops aimed at making bioinformatics more accessible to non-specialists and younger audiences.
- Fête de la Science: JeBiF hosted an interactive booth to introduce bioinformatics through engaging workshops. More than 500 visitors, both children and adults, were introduced to bioinformatics through these workshops. More precisely, workshops were focused on genome alignment with reference sequences, antibodies, DNA structure and its role. These activities aimed to explain bioinformatics and showcase its significance in modern research, to a wider audience.

The events organized by RSG France – JeBiF are open to all. Membership, which is free, provides access to our mailing list and voting rights at the general assembly.

URL

https://jebif.fr

Interplay between R-Loops and m6A RNA modification in transcriptional regulation using Drosophila S2R+ cell line

Poster

Paul Terzian [1], ***Margot Lugoboni*** [1], ***Steffen Albrecht*** [2], ***Yoan Renaud*** [1], ***Elia Ragot*** [1], ***Guillaume Junion*** [1]

1. *Université Clermont-Auvergne, IGRED,* ***2.*** *University of Auckland*

Abstract

RNAPII pausing is a step in the process of transcription regulation that was identified 50 years ago. Downstream of the transcription start site (TSS), RNAPII undergoes a brief halt in its activity, known as 'pausing'. Our team has demonstrated that the release of RNAPII is partly dependent on m6A modifications deposited on nascent RNA (1). Our work aims to reveal the molecular mechanisms that allow the release of paused RNAPII through the m6A epitranscriptomic mark. We hypothesised that another genomic structure called R-loops could be pivotal in this mechanism. We conducted a whole-genome analysis using BisMapR-seq to analyse the R-loop landscape in Drosophila and study the effect of R-loops and m6A modifiers. Here, we present the results of this analysis and show that R-loops correlate with the deposition of m6A modifications by the methyltransferase complex. We also demonstrate that the level of RNAPII pausing is increased by knocking down R-loop modulators.

URL

https://github.com/pterzian/cutandrun

InterProScan 6: a modern large-scale protein function annotation pipeline

Poster

Matthias Blum [1], ***Emma Hobbs*** [1], ***Laise Florentino*** [1], ***Alex Bateman*** [1]

1. European Molecular Biology Laboratory, European Bioinformatics Institute (EMBL-EBI)

Abstract

InterPro integrates predictive models from multiple member databases, including Pfam, to classify protein sequences into families, domains, and functional sites. InterPro annotations are widely used across the life sciences and underpin major resources such as UniProt, Ensembl, and MGnify. InterProScan is the software used to generate these annotations by scanning protein sequences against InterPro models. In addition to being distributed as a community tool, InterProScan is run in production by InterPro to annotate UniParc, and the resulting annotations are propagated to downstream resources. For example, UniProt uses InterPro annotations to support automatic functional annotation of TrEMBL entries. As Pfam annotations are now delivered through InterPro, updates to InterProScan directly affect the annotations accessed via InterPro services and APIs.

We present InterProScan 6, a complete reimplementation of the pipeline designed to support the next generation of large-scale protein annotation. The new version introduces a workflow-based architecture that improves scalability and portability across local, HPC, and cloud environments, while simplifying installation and improving reproducibility. It also enables integration of modern prediction tools, including deep learning-based methods, and supports reuse of pre-computed annotations to accelerate analyses.

Despite these major architectural changes, InterProScan 6 preserves strong continuity with previous releases. Benchmarking across diverse reference proteomes shows substantial runtime reductions, while comparisons across Swiss-Prot demonstrate near-identical annotation concordance with InterProScan 5. As the primary engine used to generate InterPro and Pfam annotations at scale, InterProScan 6 provides a modern and sustainable foundation for large-scale protein function annotation while ensuring stability for the many biological resources that depend on InterPro-derived annotations.

InterProScan 6 is distributed under the Apache 2.0 license. Its documentation is hosted on ReadTheDocs (https://interproscan-docs.readthedocs.io/en/v6/) and its source code is available on GitHub (https://github.com/ebi-pf-team/interproscan6).

URL

https://github.com/ebi-pf-team/interproscan6

INVESTIGATING GENOME REDUCTION AND EVOLUTIONARY STRATEGIES IN FRESHWATER ACTINOMYCETES

Poster

Maxime ARQUE [1], ***Gisèle BRONNER*** [1]

1. *LMGE*

Abstract

Aquatic ecosystems, both marine and freshwater, are characterized by relatively stable physicochemical conditions but strong trophic gradients that exert major selective pressures on microbial communities. In oligotrophic environments, many bacteria evolve streamlined genomes, an adaptive strategy that reduces the energetic costs of genome replication and maintenance while enhancing nutrient uptake efficiency. Genome reduction is frequently associated with metabolic auxotrophies, often explained by ecological dependencies such as those described by the Black Queen Hypothesis. However, it remains unclear whether genome streamlining in aquatic bacteria results from convergent evolutionary processes or from lineage-specific trajectories shaped by distinct selective constraints, and how these trajectories affect functional repertoires.

Here, we developed an integrative bioinformatics framework to infer ancestral functional repertoires of aquatic bacterial lineages using metagenome-assembled genomes (MAGs). This framework combines genome quality control, multi-marker phylogenomics, pangenome reconstruction, and ancestral state inference, while explicitly correcting for MAG incompleteness to minimize assembly-related biases. Ancestral genome reconstructions allowed us to quantify genome size, compactness, and reduction dynamics through evolutionary time, and to track gains, losses, and replacements of metabolic functions across phylogenies.

We applied this approach to phylogenetically related freshwater actinomycete clades, Nanopelagicales, UBA12327, and S36-B12 (family UBA5976), which share a recent common ancestor but display contrasting genomic architectures. Our analyses reveal that genome reduction in freshwater Actinomycetes follows distinct evolutionary trajectories despite shared ancestry, leading to divergent metabolic outcomes. Notably, menaquinone biosynthesis and vitamin-related pathways exhibit lineage-specific patterns of retention, loss, and reacquisition, highlighting alternative solutions to essential metabolic requirements. Together, these results demonstrate that reductive genome evolution in aquatic bacteria is not uniform but instead shaped by lineage-specific evolutionary histories, with important consequences for metabolic potential and ecological interactions.

Investigating stop codon readthrough using ribosome profiling and protein structure prediction

Poster

Enora Corler [1]

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

Abstract

Nonsense mutations generate premature termination codons (PTCs) that truncate protein synthesis and cause severe genetic diseases such as cystic fibrosis, in which ~11% of patients carry a nonsense mutation in the CFTR gene leading to a truncated protein that cannot be targeted by current modulator therapies. To restore full-length protein expression, readthrough-inducing compounds, such as aminoglycosides, offer a therapeutic strategy by promoting insertion of near-cognate tRNAs into ribosomal acceptor site. However, these molecules could also trigger off-target readthrough at natural stop codons, producing elongated proteins with unknown structural and functional consequences. Because readthrough is rare, context-dependent, and potentially disruptive, evaluating its impact is critical to ensure the safety and specificity of such therapies. This project will integrate ribosome profiling (Ribo-seq) and a protein structure prediction tool (AlphaFold) to systematically assess the consequences of induced readthrough. Multiple datasets will be analyzed to identify endogenous genes sensitive to readthrough inducers and quantify readthrough efficiency under distinct conditions. Candidate events will then be modeled with AlphaFold to evaluate how C-terminal extensions affect folding and potential function. Sequence determinants, such as stop codon motifs and 3′UTR conservation, will also be explored to uncover molecular features influencing readthrough. The integration of translation signals with structural modeling will yield a catalog of readthrough-sensitive genes, annotated with predicted elongation outcomes. This resource will support risk assessment of readthrough-inducing compounds and provide a comprehensive view of how readthrough modifies the proteome. Selected predictions will be tested through collaborations using mutagenesis or proteomic validation. Although still in progress, this work is expected to detect and validate readthrough events, reveal structural consequences of translational bypass of stop codons, and guide the development of safer, more specific readthrough-based therapies. More broadly, it will advance our understanding of translation fidelity and its modulation in health and disease.

Investigating the evolution of phototrophy in Pseudomonadota

Poster

Timothée Salzat-Hervouette [1], ***Fatoumata Mangane*** [1], ***Sophie-Carole Chobert*** [1], ***Ana Gutierrez*** [2], ***David Moreira*** [2], ***Purificación López-García*** [2], ***Fabien Pierrel*** [1], ***Sophie Abby*** [1]

1. *Univ. Grenoble Alpes, CNRS, UMR 5525, TrEE, TIMC,* ***2.*** *CNRS, Université Paris-Saclay, AgroParisTech, UMR 8079, DEEM, Laboratoire d'Ecologie Société et Evolution*

Abstract

Phototrophy is an ancient bacterial metabolism that likely originated over 3 billion years ago, prior to the rise of atmospheric oxygen provoked by oxygenic photosynthesis of Cyanobacteriia. Phototrophy enables the use of light for cellular energy production. This energy metabolism confers significant adaptive advantages to bacterial species in certain environments, yet it is sparsely distributed across several lineages of the bacterial tree of life. In the phylum Pseudomonadota (formerly Proteobacteria, also known as 'purple bacteria'), phototrophy is also sparsely distributed across various clades, yet it was proposed in the 1980s by Carl Woese that the ancestor of Pseudomonadota was a phototroph. Thus, the evolutionary origins and transmission of phototrophy in Pseudomonadota remain unresolved. The genetic potential for phototrophy is encoded by the photosynthetic gene cluster (PGC), a set of ~40 genes that colocalize in the genome and that gathers the genes required to produce photosynthetic pigments, as well as the reaction centers and the light-harvesting antennae that surround them in phototrophic chains. Previous work by others found that the PGC could be found on plasmids, and a recent study suggested that the spread of phototrophy across the family Rhodobacteraceae was facilitated by horizontal gene transfer (HGT) of plasmids containing the PGC. This study aims to determine the respective contribution of vertical and lateral transmission of phototrophy in Pseudomonadota. To this end, an annotation tool based on the MacSyFinder program was designed for the annotation of PGC in bacterial genomes. Over 2700 PGC were detected in 274 372 complete or high-quality genomes of Bacteria (mainly Pseudomonadota), which enabled subsequent phylogenomic analyses: genome context analyses, phylogenetic reconciliations and ancestral reconstructions. Overall, this study investigates the evolutionary history of phototrophy in Pseudomonadota and sets the grounds for a comparative genomic approach to investigate the metabolic factors underlying the sparse distribution of phototrophy.

KmerExploR: Fast and easy biological quality control of RNA-Seq data based on k-mers

Poster

camelia sennaoui [1], ***Chloé BESSIERE*** [1], ***Benoit GUIBERT*** [1], ***Florence RUFFLE*** [1], ***Jérôme REBOUL*** [1], ***Nicolas GILBERT*** [1], ***Thérèse COMMES*** [1], ***Anthony BOUREUX*** [1]

1. *Institute for Regenerative Medicine and Biotherapies (IRMB), U1183, Univ Montpellier, INSERM*

Abstract

KmerExploR has been first developed for inspecting large raw RNA-Seq datasets in order to improve the characterization of RNA-Seq datasets using the quantification of selected predictor genes (Riquier et al. 2021).

These genes will be used as biomarkers to define biological control in RNA-Seq datasets such as libraries strategies (polyA or ribodepletion), sequencing strategies (strand-oriented sequencing or not), contaminations sources (cell lines, virus or bacteria).

Predictor genes and their corresponding specific k-mers are first selected based on database information and the literature to address specific technical or biological questions.

Our tool is as fast as reading raw FASTQ files, and its low memory usage allows it to scale easily for very large datasets (up to several thousand samples).

In the present work, we add new biological controls as rRNA contamination or immune cell phenotypes for blood samples.

We also offer a more flexible approach with KmerExploR, enabling users to select their own collection of predictors tailored to their specific biological questions.

The tool is available at https://github.com/Transipedia/kmerexplor.

URL

https://github.com/Transipedia/kmerexplor

https://doi.org/10.1093/nargab/lqab058

Knowledge graph–driven discovery of drought tolerance genes in sorghum

Poster

Quentin SECHEr [1], ***Bill Gates Happi Happi*** [1], ***Pierre Larmande*** [1]

1. *IRD montpellier*

Abstract

Understanding the genetic basis of drought tolerance is a major challenge for crop improvement, particularly in climate-vulnerable regions. Sorghum (*Sorghum bicolor*) is a key cereal adapted to semi-arid environments, yet its functional annotation remains less comprehensive than that of model plant species such as rice, maize, and *Arabidopsis thaliana*.

To address this limitation, we developed a knowledge graph by integrating heterogeneous biological data related to sorghum, including genes, proteins, molecular interaction networks, and inter-species orthology relationships. This unified representation enables the systematic exploration of genotype–phenotype relationships across multiple biological scales.

In this work, we investigate how graph-based machine learning methods can be leveraged to predict novel gene–phenotype associations associated with drought tolerance. We explore several complementary approaches, including random walk–based algorithms, knowledge graph embedding models, and graph neural networks. In addition, orthology relationships between sorghum and well-annotated plant species (rice, maize, and *Arabidopsis*) are exploited to transfer functional knowledge across species and improve prediction performance.

The proposed methods are evaluated using known gene–trait associations derived from QTL studies, GWAS analyses, and transcriptomic datasets. Predicted candidate genes are further assessed through biological interpretation of their network context and comparison with curated plant databases such as SorghumBase.

This work demonstrates how knowledge graphs combined with graph machine learning can facilitate the discovery of candidate genes involved in complex agronomic traits, and highlights the potential of integrative data approaches for gene-trait association predictions.

Large-scale meta-omics: identifying functional signatures of marine parasitism through sequence similarity networks

Poster

Valentin Fourdraine [1], ***Clement Leboine*** [1], ***Éric Pelletier*** [1], ***Betina Porcel*** [1]

1. *Génomique Métabolique, Genoscope, Institut de Biologie François-Jacob, CEA - CNRS - Univ. Evry / Université Paris Saclay, 91000 Evry, France.*

Abstract

The ocean microbiome is a vital driver of the planet's health, yet a large part of its genetic potential remains unknown. Parasitism is a very common trophic strategy in nature and is considered a powerful evolutionary driving force for both hosts and parasites. However, marine parasites remain a black hole of biological diversity, as they are largely overlooked and underestimated. Most of these organisms are uncultivated, and the molecular tools they use to interact with their hosts are for some clades still a mystery.

This research implements a specialized bioinformatics pipeline to build a functional framework and identify molecular signatures of marine parasitism through Sequence Similarity Networks (SSN) **[1]**. In the first phase, we integrated millions of protein sequences from four major sources to cover the eukaryotic tree of life extensively: VeuPathDB, EukProt, MarFERReT, and Metagenome-Assembled Genomes (MAGs) **[2]** from *Tara* Oceans (Figure 1). We used the *DIAMOND* aligner **[3]** to perform clustering at various identity percentages to group these protein sequences into robust families.

The next stage involves the use of selected parasitic trophic-protein families to explore the *Tara* eukaryote gene **[4]** catalog, which includes over 260 million environmental unigenes. This approach would allow us to screen for candidate proteins with no prior information and to link them to potential parasitic functions through association. By cross-referencing these protein networks with physical and chemical parameters (such as temperature, nutrients, or salinity) and gene expression levels, we will attempt to better understand the adaptive strategies that marine parasites employ. This work, conducted at CEA/Genoscope, will provide a new molecular perspective on how parasites influence ocean ecosystems.

Large-scale single-cell characterization of tumor cell subpopulations in breast cancer

Poster

Quentin Rott [1], ***Odile Lecompte*** [1], ***Laurence Choulier*** [1]

1. *Université de Strasbourg*

Abstract

Precision medicine in breast cancer is currently based on a classification of tumors into three clinical and biological subtypes: hormone-dependent (HR+), HER2-positive (HER2+), and triple-negative (TNBC). Although this stratification effectively guides therapeutic decisions, treatment resistance and lack of response in some patients remain major issues, partly due to inter- and intra-tumor heterogeneity that is insufficiently captured by current diagnostic tools. Single-cell transcriptomics (scRNA-seq) technologies enable the exploration of this heterogeneity at high resolution.

In this work, a cohort of 170 samples corresponding to 130 patients from 12 independent datasets was integrated, including 100 HR+, 46 TNBC, and 14 HER2+ samples. Semi-automated pipelines were developed to preprocess the data, identify tumor cells, integrate datasets, define tumor subtypes, and characterize their transcriptional properties. Joint analysis of nearly 280,000 tumor cells identified five fundamental tumor clusters with distinct molecular signatures.

C0 does not exhibit specific enrichment relative to other clusters. C1_4_6 is enriched in pathways related to cellular stress response. C2 is associated with transcriptional programs linked to epithelial–mesenchymal transition (EMT), while C3 is enriched in proliferative pathways. C5 represents a rare population detected in most samples, expressing immune biomarkers and displaying features consistent with trogocytosed tumor cells.

Decomposition of samples according to the tumor clusters reveals a significant enrichment of EMT (C2) and proliferative (C3) clusters in TNBC, consistent with the aggressiveness of this subtype, whereas HR+ samples are enriched in clusters C0 and C1_4_6. These findings were validated in five independent datasets, as well as in spatial transcriptomics data and patient-derived xenograft models. Analysis of longitudinal HR+ cohorts resistant to therapy reveals that high proportions of C2-type cells were already present prior to treatment initiation. Taken together, our classification of tumor subpopulations deepens the understanding of breast cancer heterogeneity and provides a framework for improved prognostic stratification and treatment response prediction.

Latent Differential Graphical model for Multi-Tissue and Multi-Omics integration to model molecular interaction networks under multiple Radiation Exposure groups

Poster

Asma Nouira [1], ***Charline Jouannet*** [1], ***Maâmar Souidi*** [1], ***Catherine Ory*** [2], ***Mohamed Amine Benadjaoud*** [1]

1. *Autorité de sûreté nucléaire et de radioprotection (ASNR), PSE-SANTE/ SERAMED/Radiobiology Laboratory for Accidental Exposure,* ***2.*** *CEA DRF, iBFJ, DRCM and University Paris-Saclay*

Abstract

Exposure of the thyroid gland to ionizing radiation is associated with an increased risk of thyroid cancer, particularly papillary thyroid carcinoma. While the carcinogenic effects of high-dose radiation are well established, the biological consequences of low-dose exposure remain poorly understood due to limited statistical power, long latency periods, and the absence of specific pathological markers. The availability of multi-omics measurements from the Chernobyl cohort, including transcriptomic and miRNA expression profiles obtained from normal and tumoral thyroid tissues, provides an opportunity to investigate radiation-associated molecular alterations through integrative network analysis. In this work, we propose a latent differential graphical model for the joint analysis of multi-tissue and multi-omics data across multiple radiation exposure groups. The proposed framework integrates gene and miRNA expression measured in normal and tumor tissues by constructing a concatenated multi-omics representation for each exposure group. To capture shared biological patterns across tissues and molecular layers within individuals, we introduce a latent Gaussian graphical model incorporating a low-rank component. The baseline molecular interaction network is modeled through a sparse plus low-rank decomposition, while deviations from this baseline are captured by sparse differential networks corresponding to low- and high-dose radiation exposure. Model parameters are estimated through penalized likelihood optimization that enforces sparsity in conditional dependencies and low-rank latent structure. A fusion penalty is introduced to encourage similarity between exposure groups while allowing dose-specific perturbations. Simulation studies under multiple scenarios demonstrate that the proposed method improves precision and recall in recovering network structures and differential interactions compared with existing approaches. Application to the Chernobyl cohort reveals dose-dependent molecular network reorganization, enabling the identification of radiation-specific and dose-specific molecular interactions. In addition, tumor networks appear denser than those observed in normal tissues.

These results highlight the ability of our model to reveal molecular mechanisms underlying dose-dependent radiation effects in human thyroid gland.

URL

Software:

https://github.com/asmanouira/Multi-DiffNet

Leveraging atlas-level single cell resources as reference panels for bulk RNA-seq deconvolution.

Poster

Martina Gallinaro [1], Marie-Laure Plissonnier [2], Armando Andres Roca Suarez [2], Giovanni Malerba [1], Massimo Levrero [3], Massimiliano Cocca [2]

1. *Biology and Genetics Section, Department of Neurosciences, Biomedicine and Movement Sciences, University of Verona,* ***2.*** *UMR PaThLiv Inserm 1350 Université Lyon 1 (UCBL1) - The Lyon Hepatology Institute EVEREST,* ***3.*** *UMR PaThLiv Inserm 1350 Université Lyon 1 (UCBL1) - The Lyon Hepatology Institute EVEREST; Department of Hepatology, Croix Rousse hospital, Hospices Civils de Lyon*

Abstract

Tissue heterogeneity and cell-type abundances critically influence physiology and can obscure disease mechanisms in bulk molecular analyses. While single-cell RNA-seq (scRNA-seq) provides high resolution, its cost and technical limitations make computational deconvolution an appealing alternative for inferring cellular compositions from bulk transcriptomics. Nonetheless,, deconvolution performance is often hampered by inconsistent annotations and the limited generalizability of study-specific reference panels.

To address these challenges, we developed a standardized workflow to construct a liver-specific, atlas-level reference panel. We integrated four datasets (GSE149614, GSE247128, GSE136103, GSE202379) encompassing healthy and pathological conditions, including Hepatocellular Carcinoma (HCC), cirrhosis, and steatosis. Integration was performed via Seurat/Harmony, followed by an automated, marker-based annotation using ScType and curated signatures from Cell Marker 2.0 and Azimuth. The final resource comprises 188,727 cells across 45 samples, structured into three levels of granular annotation.

We benchmarked the atlas using simulated pseudo-bulk and real Precision Cut Liver Slices (PCLS) data, comparing state-of-the-art tools: MuSiC2, CDSeq, CIBERSORTx, and a VAE-based method (Bulk2Space). Our results demonstrate that this atlas-level reference improves the accuracy of cell-type inference compared to narrow, “perturbed” reference sets. We further applied this resource to a clinical cohort of 57 MASH-related HCC patients (paired tumor/non-tumor), successfully reconstructing cellular heterogeneity in a clinically relevant context.

Our workflow provides a robust, standardized framework for dataset integration and annotation, to create of a useful and extensible resource. Future work will explore advanced deep learning architectures, such as Domain Invariant Variational Autoencoders (DIVA), to better capture the complex variability of bulk RNA-seq data and improve the generalizability of the resource across diverse experimental conditions.

LLM Training Dataset for Plant Biology & Food Processing Literature

Poster

Tom Colombu [1], ***Guillaume Laisney*** [2], ***Clément Frainay*** [2], ***Franck Giacomoni*** [3], ***Magalie Weber*** [4], ***Olivier Filangi*** [1]

1. *Institute for Genetics, Environment and Plant Protection (IGEPP), National Research Institute for Agriculture, Food and Environment (INRAE), Institut Agro, Université Rennes,* ***2.*** *Toxalim (Research Center in Food Toxicology), Université de Toulouse, INRAE, ENVT, INP-Purpan, UPS,* ***3.*** *University Clermont Auvergne, INRAE, UNH, Metabolism Exploration Platform, MetaboHUB Clermont,* ***4.*** *INRAE, UR BIA (Biopolymers Interactions Assemblies)*

Abstract

The scientific literature abounds with information highlighting the central role of numerous metabolites in plant-derived products, whether in disease resistance, environmental interactions, or organoleptic properties. Such knowledge is essential for understanding the mechanisms underlying plant traits, their responses to biotic and abiotic stresses, and their suitability for processing.

However, despite this wealth of information, the effective reuse of existing knowledge in research projects remains challenging. In contrast to the biomedical field—where data standardisation and knowledge indexing are more advanced—the plant sciences still suffer from fragmented information, hindering the integration and exploitation of available data.

In this context, the Semantic Metabolomics Data Lake, a Big Data infrastructure designed to generate and exploit knowledge graphs that contextualise data from metabolomics platforms, incorporates natural language processing (NLP) components. These components rely on state-of-the-art artificial intelligence (AI) methods to automatically annotate the entire body of scientific literature using reference ontologies from the plant domain, such as the Plant Ontology (PO), Trait Ontology (TO), and Plant Experimental Conditions Ontology (PECO) developed within the Planteome project, as well as the TransformON ontology dedicated to food transformation processes.

Recent approaches based on language models (LLMs) built on transformer architectures have enabled more reliable text annotation. Their ability to assess semantic similarity allows for efficient alignment between ontological term definitions and scientific text content.

This work presents the development and publication of a training dataset designed to enhance the annotation and contextualisation of scientific literature in plant metabolomics. The ultimate goal is to specialise a language model in semantic comparison between scientific abstracts and ontology terms, thereby improving the accuracy of the annotation processes integrated into the Semantic Metabolomics Data Lake.

URL

https://hal.inrae.fr/hal-05075616v1

madbot national working group : join us to participate to the development and adoption of madbot for FAIR data and metadata management

Poster

Imane MESSAK [1], ***Baptiste Rousseau*** [1], ***Elora Vigo*** [1], ***madbot working group*** [1], ***Hélène CHIAPELLO*** [1], ***Nadia Goué*** [2], ***Julien Seiler*** [1], ***Thomas Denecker*** [1]

1. *IFB-core, Institut Français de Bioinformatique (IFB), CNRS, INSERM, INRAE, CEA,* ***2.*** *Plateforme AuBi, Mésocentre, UCA*

Abstract

In the context of Open Science and increasingly intensive data production, researchers are expected to manage and share their data according to the FAIR principles. However, implementing FAIR-compliant data and metadata practices remains challenging and time-consuming for many teams. Technical complexity, heterogeneous standards, and limited support resources create bottlenecks, highlighting the need for dedicated tools and coordinated support structures.

To address these challenges, madbot (Metadata and Data Brokering Online Tool) was developed by Institut Français de Bioinformatique, as a platform designed to facilitate FAIR-aligned data and metadata management. To ensure madbot is developed based on user needs, its deployment and adoption are supported by a multidisciplinary national working group bringing together scientists, data stewards and engineers distributed across France. Working alongside the core development team, this group follows and contributes to the evolution of the platform through user-oriented activities including functional testing, feedback collection, and evaluation of real-world use cases across diverse scientific domains. By regularly reviewing new features and participating in interactive testing phases, this group ensures that madbot remains aligned with community needs and practical research workflows.

The group also plays an important role in disseminating madbot within the biology research community by contributing to training sessions, demonstrations, and outreach initiatives that encourage the adoption of FAIR data management practices. Its geographically distributed structure strengthens connections with regional platforms and research teams, multiple institutions and scientific communities across France, facilitating broader awareness and uptake. Acting as a bridge between developers and user communities, the working group supports continuous improvement, dissemination, and long-term adoption of madbot within the open science ecosystem.

This contribution presents the operational framework and lessons learned from this national deployment, highlighting how a distributed working group can foster sustainable adoption of FAIR-oriented tools. This group is completely open, and anyone wishing to join is welcome.

URL

https://madbot.france-bioinformatique.fr/

MetaPanG: a pangenome graph-based method for strain-level profiling of prokaryotic microbiomes

Poster

Téo Lemane [1], ***Jean Mainguy*** [2], ***Claudine Médigue*** [1], ***Alexandra CALTEAU*** [2], ***David VALLENET*** [2]

1. *1. LABGeM, Génomique Métabolique, CEA, Genoscope, Institut François Jacob, Université d'Évry, Université Paris-Saclay, CNRS, FRANCE 2. Institut Français de Bioinformatique, IFB-core, UAR 3601, CNRS, Villejuif, France,* ***2.*** *LABGeM, Génomique Métabolique, Genoscope, Institut François Jacob, CEA, CNRS, Université Évry, Université Paris-Saclay*

Abstract

Metagenomic taxonomic classification is essential for understanding microbial community composition. Most existing tools rely on gene markers or flat reference genomes and provide limited functional context beyond species identification. By relying on gene-centric pangenome graphs (1), we can leverage intraspecific diversity and graph structure to achieve strain-level profiling while simultaneously providing gene content of each strain for further functional characterization. Here, we present MetaPanG, a new tool for metagenomic reads profiling that uses the pangenome collections hosted by PanGBank as a reference database (2). MetaPanG aims to identify and quantify strains present in a metagenomic sample, along with their associated gene families, while highlighting genes that may be undetected due to limited sequencing depth.

The profiling process proceeds in three stages. First, a sourmash (3) signature is computed from the metagenomic reads and compared against the indexed signatures of PanGBank genomes to estimate the coverage of each species by the sample. This step allows rapid identification of species candidates likely represented in the mix. Second, for each identified species, reads are assigned to gene families of the corresponding pangenome by k-mer matching against a precomputed pangenome de Bruijn graph annotated with gene family identifiers using MetaGraph (4). It produces a weighted pangenome graph reflecting gene family abundances. Third, strain identification and relative abundance are inferred by solving a non-negative least squares problem on the weighted graph, followed by determination of the gene family content of each strain.

To ensure scalability, all resource-intensive preprocessing, such as signature and de Bruijn graph construction, is performed once per pangenome and stored within PanGBank. All required resources are retrieved on-demand through the PanGBank API during the profiling process.

MetaPanG is currently under active development. We will present benchmarking results on synthetic metagenomic datasets exhibiting the feasibility and value of pangenome-aware taxonomic profiling at strain-level.

Methodological approach for RNA edition analysis: a brain tissue case study

Poster

Julie Le Borgne [1], ***Florence Mauger*** [1], ***Marie Bouaud*** [1], ***Florence Jobard*** [1], ***Christian Daviaud*** [1], ***Bertrand Fin*** [1], ***Francis Rousseau*** [1], ***Eric Bonnet*** [1], ***Jean-François Deleuze*** [1], ***Kévin Muret*** [1]

1. *Centre National de Recherche en Génomique Humaine (CNRGH), IBFJ, CEA, Université de Paris-Saclay, Evry, France*

Abstract

Despite the identification of several vital RNA editing sites in humans, their role remains little known, due to a high prevalence of false positive detection. While some sequencing-based approaches have sought to identify RDD (RNA-DNA difference site), they remain insufficiently validated. Thus, using data from 33 brain samples in cerebellum or hippocampus from 19 unrelated individuals on a transcriptome-wide scale, we focused on developing a rigorous method to effectively filter out false positives and improve reliability of RDD detection. First, for each sample, we generated WGS (30x) and two stranded deep coverage RNA sequencing (~140M paired-end reads) with UMIs. Then, our pipeline based on JACUSA2 software integrates: 1) a strict RNA read mapping and processing, namely by excluding PCR duplicates and multimapped reads, 2) the exclusion of low-complexity regions; we considered only 'easy' regions (Heng Li. 2025 *Gigascience*) and excluded loci near indels or splice sites, which are prone to misalignments, 3) the application of RDD filters, such as end-of-read bias or unidirectional bias, and 4) the reproducibility of results with technical duplicates. With this method, we identified 468,096 robust RDDs, 232,919 of which were found in at least three samples in either tissue. Among them, 98.7% were, as expected, A-to-G (corresponding to A-to-I), the most prevalent RNA edition in human, mediated by ADARs. Additionally, 95.7% of those RDDs were identified in the literature and 92.0% were in ALU regions, which are expected as most known RDDs were in ALU and most adenosines in ALU are subject to edition (Lily Bazak et al. 2014 *Genome Res*.). Finally, 4.7% and 18.2% RDDs were in protein coding gene exon and UTR, respectively; such editing may lead to amino-acid substitution or alter gene expression. Next, we will investigate the variation of RNA editing level between tissues and individuals.

mETHYLotest: a unified toolkit for multi-platform DNA methylation analysis

Poster

Nicolas Doldi [1], ***Maud De Dieuleveult*** [1], ***Patrick Nitschke*** [1], ***Emilia Puig Lombardi*** [1]

1. *Institut Imagine, Université Paris Cité, INSERM U1163*

Abstract

DNA methylation is a key epigenetic mark involved in gene regulation, development, and disease. Its analysis relies on a wide array of technologies, from Illumina Infinium MethylationEPIC arrays to short-read and long-read sequencing approaches such as whole-genome bisulfite sequencing (WGBS) or nanopore sequencing. This technical heterogeneity typically requires researchers to use disparate R packages, each with its own data structures, conventions, and interfaces, complicating the implementation of unified and reproducible analytical workflows.

To address this challenge, we developed mETHYLotest, an R library that provides a streamlined and modular framework for DNA methylation analysis. Rather than imposing a single data structure, mETHYLotest offers two dedicated and optimized pipelines: one powered by ChAMP for array-based data, and another by MethylKit for sequencing data. The package features automated quality control checks, integrated analysis pipelines, and intuitive user interfaces (UI), while simplifying deployment and installation. Built upon the native R objects of its core engines, mETHYLotest ensures full interoperability, allowing researchers to seamlessly integrate these tools into their existing workflows.

Developed within the Bioinformatics Platform of the Imagine Institute (Paris), mETHYLotest is integrated into the facility's core analytical pipelines. This institutional integration guarantees professional-grade maintenance and long-term support. Adhering to FAIR principles, mETHYLotest is freely available as an open-source R package, ensuring transparency and reusability for the epigenomics community.

The GitHub repository is now publicly available at https://github.com/OMICShub-Imagine/mETHYLotest, and the manuscript is currently under review.

URL

https://github.com/OMICShub-Imagine/mETHYLotest

Mfd's connectors at the heart of its extensive reshaping

Poster

Thomas Marino [1], ***Samantha Samson*** [1], ***Sylvain Marthey*** [1], ***Nalini Rama Rao*** [1], ***Gwen Andre*** [1]

1. *INRAE*

Abstract

Amid the urgent global crisis of antimicrobial resistance (AMR), the bacterial protein Mfd has emerged as an innovative drug target. Mfd is a key virulence factor that repairs DNA damage, caused by nitric oxide (NO) produced by the macrophages during infection. As a non-essential transcription-repair coupling factor unique to bacteria, Mfd recognizes RNA polymerase stalled at non-coding lesions [1]. Using ATP fuel, Mfd is able to disassemble the transcription complex and to recruit the nucleotide excision repair machinery, thereby promoting bacterial survival and mutation. In the fight against antibiotic resistance, Mfd represents a promising therapeutic target that must be neutralized. Our first objective was to inhibit its ATPase activity.

In a previous study, we identified NM102 as a competitive inhibitor that blocks the ATP-binding site of Mfd [2]. Our second objective is to interfere with the extensive three- dimensional remodeling that Mfd undergoes to engage in its multiple functional partnerships. Combining sequence evolution analyses, molecular dynamics simulations, and targeted substitutions, we are characterizing the dynamic patterns that drive remodeling along key conformational states revealed by cryo-electron microscopy [3]. Notably, we have identified two long linkers flanking the central RNA polymerase–binding domain that may couple their motions to orchestrate the global remodeling of Mfd. During the poster session, I will present our ongoing work.

Microbial transfers in dairy compartments under two farming systems

Poster

Sébastien Theil [1]***, Mahendra Mariadassou*** [2]***, Philippe Ruiz*** [3]***, Guillaume Kon Kam King*** [2]***, Céline Delbès*** [1]***, Anne-Laure Abraham*** [2]

1. *UMR 545 Fromage, Université Clermont Auvergne, INRAE, VetAgro Sup, Aurillac, France,* ***2.*** *Université Paris-Saclay, INRAE, MaIAGE, 78350, Jouy-en-Josas, France; Université Paris-Saclay, INRAE,* ***3.*** *Université de Toulouse, INRAE, UR 875 MIAT, F-31320, Castanet-Tolosan, France*

Abstract

Farming practices affect the microbial characteristics of the food matrix, particularly in raw milk cheese production. To address this issue, we explored microbial transfers across cheese production from farm environments to raw milk cheese and rat microbiota consuming these cheeses. Cows were placed in two farming systems — agroecological (AE) and intensive (IN) — with varying proportions of grazing in their diet for 3 months. In order to identify microbial transfers, 340 samples were collected in 12 compartments (air, soil, grass, cow rumen and feces, teats, bedding material, milk, milk filter, cheese curd, cheese rind, and rat feces) at three sampling times. A threefold approach was used to (1) characterise the microbial diversity of the twelve compartments, (2) identify microbial fluxes between compartments, and (3) assess the impact of the farming system and time on both the diversity and the fluxes.

The bacterial and fungal communities were characterized by 16S V3-V4 and ITS sequencing. Data processing was performed using the rAnomaly workflow. The twelve compartments exhibit substantial differences in community composition and diversity. As expected, beta diversity analyses showed that the compartment was the main structuring factor. We then analysed all compartments separately and identified genera differentially abundant between the two farming systems (DESeq2 analysis). We finally focused on microbial transfers between compartments. We constructed a network of shared ASVs across the farm-to-fork agri-food chain. The network exhibited diversity hotspots and bottlenecks affecting downstream microbiota, including bedding material and the cow's rumen. Finally, we used a source - sink approach, using a modified version of the FEAST algorithm, to quantify the contribution of nearby environments to the milk microbiota with uncertainty quantification. These results will provide more insight into fluxes across a representative agri-food chain, and the methodological approach and analytical tools can be transferred to other ecosystems.

MicroScope, an Integrated Platform for the Annotation and Exploration of Microbial Gene Functions through Genomic, Pangenomic and Metabolic Comparative Analysis

Poster

Alexandra CALTEAU [1], ***Noëlle Haddad*** [1], ***Aurélie Lajus*** [1], ***Jean Mainguy*** [1], ***David Roche*** [1], ***Zoé Rouy*** [1], ***David VALLENET*** [1]

1. *LABGeM, Génomique Métabolique, Genoscope, Institut François Jacob, CEA, CNRS, Université Évry, Université Paris-Saclay*

Abstract

Large-scale genome sequencing and high-throughput approaches generate vast data, revolutionizing our understanding of thousands of microbial species. Yet, fully interpreting these genomes remains a challenging task for microbiologists. To address this challenge, we develop MicroScope, an integrated Web platform for management, annotation, comparative analysis and visualization of microbial genomes (https://mage.genoscope.cns.fr/microscope) [1]. The platform enables collaborative work in a rich comparative genomic context and improves community-based curation efforts.

Launched in 2005, MicroScope provides analyses for complete and ongoing genome projects together with metabolic network reconstruction and transcriptomic experiments allowing users to improve the understanding of gene functions. Besides automatic functional annotations, we integrated several tools to analyze a wide range of biological systems (antibiotic resistance, secondary metabolites, defense systems,...). Particularly, tools from the PPanGGOLiN software suite allows users to analyze pangenomes from several hundreds of genomes of the same species and to explore their content in regions of genomic plasticity [2,3]. The platform also has extensive functionality to explore and compare metabolic pathways.

MicroScope is widely used by microbiologists from academia and industry all around the world for collaborative studies and expert annotation. To date, MicroScope contains data for ~25,000 microbial genomes, part of which are manually curated and maintained by microbiologists (>7,500 user accounts in March 2026 among which only 35% are from France). We offer professional training, but the platform is also a useful resource for academic training.

References

1. Vallenet D., Calteau A., et al. MicroScope: an integrated platform for the annotation and exploration of microbial gene functions through genomic, pangenomic and metabolic comparative analysis. Nucleic Acids Research, Jan 8;48(D1):D579-D589, 2020.
2. Gautreau G. et al. PPanGGOLiN: Depicting microbial diversity via a partitioned pangenome graph. PLoS Computational Biology, Mar19;16(3):e1007732, 2020.
3. Bazin A. et al. panRGP: a pangenome-based method to predict genomic islands and explore their diversity. Bioinformatics, Dec 30;36(Suppl_2):i651-i658, 2020.

URL

https://mage.genoscope.cns.fr/microscope

MIMEco: Multi-objective metabolic modeling to predict and explain pairwise ecosystem interactions

Poster

Anna Lambert [1], ***Samuel Chaffron*** [2], ***Damien Eveillard*** [3]

1. *Institute for Quantitative and Computational Biosciences (IQCB), Johannes Gutenberg-University Mainz, Mainz, Germany,* ***2.*** *Nantes Université, École Centrale Nantes, CNRS, LS2N, UMR 6004,* ***3.*** *Nantes Université*

Abstract

Predicting microbial interactions is essential for understanding ecosystem dynamics, such as the soil, marine, or gut communities. GEnome-scale metabolic Modeling (GEM) approaches allow for a mechanistic exploration of the metabolic interactions in these ecosystems , but they rely on community-wide objective functions (misrepresenting the group of self-prioritizing organisms that is a real ecosystem) or sample-specific abundance data (dependant on a specific observed condition). We present MIMEco (Metabolic Interaction Modeling in Ecosystems), an open-source Python package that utilizes formal multi-objective optimization to predict “egoistic” pairwise microbial interactions. MIMEco identifies interaction types, quantifies interaction potential and identifies exchanged metabolites directly from Pareto fronts without requiring abundance data. We validated MIMEco using an in vitro co-culture of auxotrophic Escherichia coli strains, accurately replicating growth recovery and syntrophic cross-feeding.

URL

Preprint: https://www.biorxiv.org/content/10.1101/2025.11.25.690486v1.full.pdf
Github: https://github.com/Anna-cell/mimeco

Minimal feature set selection for spatial transcriptomics data clustering and preventing over-clustering

Poster

Tess Chilliet [1], ***Christophe Le Priol*** [1]

1. *Université Paris Cité, CNRS, Inserm, Institut Cochin*

Abstract

In a classical transcriptomics data clustering workflow, valuable genes to use for clustering are selected after the normalization of raw sequencing counts. The widely used Seurat package selects highly variable genes (HVG) genes. However, other feature selection methods, based on highly expressed (HEG) and highly deviated genes (HDG), have been shown to outperform the HVG method. Besides, selecting anti-correlated genes has been shown to prevent over-clustering, i.e. the erroneous partitioning of homogeneous datasets. Here, we simulated spatial transcriptomics data and evaluated the performance of these methods in identifying valuable genes for spot clustering. We evaluated decreasing feature set sizes in order to identify minimal gene sets that generate high-quality clusterings while avoiding over-clustering. Overall, the HVG method is systematically outperformed by the HEG and HDG methods and fails to correctly cluster the data. On the contrary, HEG and HDG perfectly cluster the data for a high fold-change (2) and 50 to 150 selected features. Gene expression correlation heatmap can be used to assess the relevance of selected features for a clustering purpose. If signals of both clusters are correctly picked, two sets of anti-correlated features will appear on the map. Since two clusters are simulated, we identified three sets of features based on the correlation heatmap. The sets that are valuable for spot clustering can be easily identified thanks to their anti-correlation profile, whereas uninformative features have a null mean correlation. We showed that the identification of relatively small sets of anti-correlated features with the HEG and HDG methods may generate good-quality clusterings and prevent over-clustering.

Modernizing ATGC Bioinformatics Services: Migration to a Shared Meso-Centre and API-Driven Delivery

Poster

Christophe Menichelli [1], ***Sylvain Milanesi*** [2], ***Stéphane Guindon*** [1], ***Eric Rivals*** [1], ***Laurent Bréhélin*** [1]

1. *LIRMM, Université de Montpellier, CNRS,* ***2.*** *LIRMM, Université de Montpellier, CNRS, Institut Français de Bioinformatique*

Abstract

ATGC supports bioinformatics research through software dissemination, online analyses, and project-oriented expertise in phylogeny, comparative genomics, and NGS data analysis. In 2024, the platform served about 12,000 distinct users and handled 370,000 analyses, totaling around 160,000 CPU hours. This growing usage highlighted the need to redesign the infrastructure to improve maintainability and scalability.

In 2025, ATGC launched a modernization effort addressing heterogeneous legacy components, fragmented access points, and limited operational scalability. Services were migrated to the regional meso-centre (cluster IO) to mutualize compute and storage resources while simplifying service delivery for both users and developers.

The new architecture separates access, service logic, and execution layers. A WordPress front-end provides tool discovery, job submission, and result visualization. A Django/DRF API called TIDE, designed as a successor to WAVES (Chakiachvili et al., 2019), exposes a unified catalog of tools, orchestrates jobs, and enables programmable access to services and metadata. Job execution is handled by a shared SLURM layer running on mutualized CPU/GPU resources.

A key design choice was to maintain a unified catalog for both online and offline tools. Online tools are fully integrated in TIDE, while offline tools remain accessible through curated software pages and can be progressively integrated into online execution when requirements are met. This dual-mode approach enables gradual modernization without disrupting user habits.

To improve reproducibility and facilitate developer onboarding, we defined a standardized integration workflow based on EDAM metadata (Ison et al., 2013), Apptainer containers, and well-defined execution interfaces.

This migration illustrates how a medium-sized bioinformatics platform can combine local scientific support, shared HPC infrastructure, and reproducible service engineering to improve long-term sustainability.

MSEABOARD: An open source and web-based interactive platform for linked visualization and analysis of bioinformatics data.

Poster

Luca Nesterenko [1]

1. *PRABI, Rhône-Alpes Bioinformatics Center, Université Lyon 1, 43 bd du 11 novembre 1918, Villeurbanne CEDEX 69622, France*

Abstract

Interactive exploration of multiple sequence alignments (MSAs), phylogenetic trees, and structural data is central to molecular biology, evolutionary genomics, and structural bioinformatics. Existing tools provide powerful capabilities but typically only focus on a single data modality, leading researchers to juggle between several programs.

MSEABOARD is an open source and highly flexible web-based visualization tool for different types of biological data. It supports and has a comprehensive set of features for displaying several data types: genetic sequences, multiple sequence alignments, phylogenetic trees (with support for node annotations), sequence logos, distance matrices, general tabular data and protein structures.

With a customizable layout, the tool enables visualizing different data types in different panels which can be further linked to each other allowing for interactive and interconnected visualizations, for instance an MSA can be linked to a corresponding phylogenetic tree: When hovering over a sequence in the MSA panel the corresponding leaf in the tree will be highlighted and vice-versa. Likewise, a distance matrix can be linked to an MSA, a tree or a protein structure, hovering over a cell in the matrix will then highlight the two corresponding sequences/tree nodes/residues in the 3D structure.

Beyond visualization, the platform allows a considerable degree of data manipulation and analysis, users can for instance create a sequence logo or per-site statistics from an MSA, a distance matrix from an MSA, a tree or a protein structure. Some lightweight tree reconstruction features are implemented as well.

The stated goal of the project is to bridge the gap between biological data and visualizations thereof, via a single and easily accessible tool instead of a fragmented ecosystem, making beautiful, interactive and interconnected visualizations readily available for bioinformatics researchers and students. These features can speed up workflows and help gain insight into the data by bringing them to life.

URL

https://www.mseaboard.com/

Multi-omics network integration across disease progression in myotubular myopathy

Poster

Supriya Priyadarshani SWAIN [1], ***Anaïs Baudot*** [2], ***Jocelyn LAPORTE*** [1]

1. *IGBMC,* ***2.*** *CNRS*

Abstract

X-linked myotubular myopathy (*XLMTM*) is a rare and severe form of centronuclear myopathy (CNM) caused by loss-of-function mutations in Myotubularin 1 (*MTM1*). Previous studies have used network-based integration to combine different layers of omics with public knowledge bases into a multilayer network, revealing both pathogenic and protective pathways in XLMTM. However, the analyses overlooked the temporal dynamics of disease progression across developmental stages. Moreover, the coordinated impact of transcriptomic and proteomic changes on downstream metabolic alterations remains poorly understood.

To address this, we performed longitudinal transcriptomic and proteomic analyses in the tibialis anterior muscle of *Mtm1*$^{-/y}$ mice at embryonic, early, and late developmental stages. These analyses revealed temporal dysregulation at the pathway level, with coordinated changes in gene and protein expression across stages. Notably, pathway-level overlap between transcriptomic and proteomic layers highlighted convergent molecular alterations, emphasizing the need for time-resolved integrative approaches to capture disease progression.

Building on these findings, we propose a temporal multi-omics network framework that integrates transcriptomic, proteomic, metabolomic, and lipidomic datasets across developmental stages. Transcriptomic and proteomic profiles are modelled as a time-resolved multiplex network, in which each stage represents a distinct layer and is connected by directed temporal edges. Metabolomic and lipidomic data collected at the late stage are incorporated via pathway-based bipartite connections linking genes, proteins, and metabolites. Network propagation using Random Walk with Restart (MultiXRank), seeded from *Mtm1* in the early stage, enables the topological and functional prioritization of molecular features and pathways and potentially captures the dynamic disease trajectories .

Overall, this provides a comprehensive view of disease progression, facilitates identification of candidate biomarkers, and highlights potential therapeutic targets for XLMTM.

Multi-reference STARR-seq analysis reveals candidate enhancers associated with the 2La inversion in Anopheles

Poster

Adrien Pain [1]

1. *Institut Pasteur, Université Paris Cité, Bioinformatics and Biostatistics Hub, Paris, France*

Abstract

Chromosomal inversions are structural variants that can contribute to ecological adaptation by maintaining linked combinations of alleles. In the malaria mosquito species complex including Anopheles gambiae, the polymorphic 2La inversion spans ~20 Mb on chromosome 2L and has been associated with multiple adaptive traits, including mosquito resting and biting behavior, ecological adaptation, insecticide resistance, and variation in malaria transmission capacity. While the evolutionary significance of this inversion has been extensively studied, the regulatory mechanisms underlying phenotypic differences between its alternative haplotypes remain poorly understood.

Transcriptional enhancers are key regulators of gene expression and can influence multiple target genes. Recent STARR-seq experiments have enabled the genome-wide identification of enhancer elements in Anopheles. However, functional genomics analyses are typically performed using a single reference genome, which may introduce reference bias when analyzing structurally polymorphic regions such as chromosomal inversions.

To investigate the regulatory landscape of the 2La inversion while accounting for potential reference bias, we reanalyzed STARR-seq data generated from wild mosquitoes carrying the 2La haplotype using multiple reference genomes representing alternative inversion configurations. Independent enhancer catalogs were generated for each reference genome and subsequently compared to identify candidate enhancers associated with the 2La haplotype. This multi-reference approach revealed a set of enhancers that were not detectable when using a single reference genome, highlighting how reference choice can influence enhancer discovery in structurally complex regions. Functional assays performed on a subset of candidates confirmed differences in enhancer activity between haplotypes.

Together, our results reveal regulatory differences between 2La and 2L+ haplotypes and highlight the value of multi-reference approaches for functional genomics analyses in structurally variable genomic regions.

NARCOD: Non-Arbitrarily Reproducible Clustering of transcriptOmics Data

Poster

Maryline Favier [1]***, Rachel Onifarasoaniaina*** [1]***, Hélène Collinot*** [2]***, Djihane Djeridane*** [2]***, Tess Chilliet*** [2]***, Sébastien Jacques*** [3]***, Daniel Vaiman*** [2]***, Céline Méhats*** [2]***, Christophe Le Priol*** [2]

1. *Université Paris Cité, CNRS, Inserm, Institut Cochin, Histim,* ***2.*** *Université Paris Cité, CNRS, Inserm, Institut Cochin,* ***3.*** *Université Paris Cité, CNRS, Inserm, Institut Cochin, Genom'IC*

Abstract

Clustering samples with similar transcriptomic profiles is one of the first steps in the analysis of omics data. The algorithms used for this task are affected by randomness. To achieve reproducibility of the outcome, the seed, which controls the random number generator, must be set to a constant value. However, this arbitrary way to achieve reproducibility is rarely addressed in biological studies.

Here, we showed that the outcome of the classical clustering approach, which consists in analysing an entire dataset in a single round of clustering, can not generate non-arbitrarily reproducible clusterings, i.e. without setting the seed to a constant value. We developed a new method, NARCOD (Non-Arbitrarily Reproducible Clustering of transcriptOmics Data), to make the clustering of omics data insensitive to algorithm randomness. We showed that this objective can be achieved by breaking down the challenging task of clustering a diverse high-dimensional dataset into successive simple clusterings via a recursive process.

We applied our method and the classical clustering approach to multiple 10x Genomics Visium datasets. Unlike the classical approach, we showed that our method can generate non-arbitrarily reproducible clusterings while maintaining biological relevance.

Neuronal epigenetic plasticity in polyaddictions

Poster

Yahia Hadj-Arab [1], ***Esther Colantonio*** [2], ***Margot Diringer*** [3], ***Mathieu Bruggeman*** [3], ***Emmanuel Darcq*** [2], ***Anaïs Bardet*** [1], ***Pierre-Eric Lutz*** [3]

1. *IGBMC,* ***2.*** *CRBS,* ***3.*** *INCI*

Abstract

Substance use disorders (SUD) are a group of chronic psychiatric disorders characterized by compulsive use of psychoactive substances despite harmful consequences. Although these substances have distinct molecular targets (opioid, cannabinoid, nicotinic receptors, etc.), they all activate the dopaminergic signaling targeting the prefrontal cortex, a common mechanism at the core of pathophysiology.

Despite this convergence, most molecular studies of SUD in the human brain focused on a single drug of abuse and cohorts of modest size mostly limited to male subjects, using methods that did not interrogate the full genome.

To address these limitations, during my PhD project I am conducting multiomic, genome-wide analyses of prefrontal cortex tissue from a large cohort of patients with SUD and controls (n=114 total). To characterize the epigenetic plasticity implicated in SUD, we analyzed DNA methylation at cell-type specific level (focusing on neuronal nuclei isolated by flow cytometry), as well as its functional impact on the transcriptome. In addition, based on our recent preclinical work suggesting significant sex differences in epigenetic effects of drugs of abuse, our study stands out by the systematic inclusion and analysis of individuals of both sexes.

Overall, our results should provide new knowledge related to sex-specific epigenetic mechanisms of addiction.

ONTmethPLANT: a reproducible pipeline for integrated analysis of DNA methylation and genomic variants from Oxford Nanopore data in plants

Poster

Mame Seynabou FALL [1]

1. *INRAE – IPS2 (Institut des Sciences des Plantes Paris-Saclay)*

Abstract

DNA methylation is a major epigenetic modification involved in gene regulation and genome stability in plants. Oxford Nanopore Technologies (ONT) long-read sequencing enables the simultaneous characterization of DNA methylation and genomic variation from the same sequencing dataset. However, extracting these complementary layers requires robust, reproducible, and easily deployable bioinformatics workflows.

Here, we present ONTmethPLANT, a modular pipeline implemented using Snakemake for the integrated analysis of ONT sequencing data. Starting from raw signal files (e.g. pod5), the workflow supports both barcoded and non-barcoded datasets and enables users to perform methylation analysis, variant analysis or both. The two modules share core preprocessing steps, including basecalling, read alignment, and the use of genome annotation files for downstream analysis. The methylation module performs methylation calling and identifies differentially methylated regions (DMRs), whereas the variant module detects SNPs and structural variants (SVs).

To ensure portability and reproducibility, ONTmethPLANT is distributed within a Singularity-based container, enabling straightforward deployment across high-performance computing infrastructures. The pipeline produces standardized outputs, including genome-wide methylation profiles, DMRs, SNPs and SVs with their annotation, and variant call files. Overall, ONTmethPLANT provides a scalable and reproducible framework for the joint analysis of epigenetic and genetic variation from long-read sequencing data in plants.

URL

Software repository (restricted access): https://forge.inrae.fr/sps-bioinfo/ontmethplant

Open Science: A Catalogue of European Tools Supporting Research Data Management

Poster

Saliha Zenboudji-Beddek [1], ***Jean-François Dufayard*** [2], ***Sylvain Milanesi*** [3], ***Christophe Bruley*** [4], ***Anne-Françoise Adam-Blandon*** [5]

1. *IFB-core, Institut Français de Bioinformatique (IFB), CNRS, INSERM, INRAE, CEA, 94800 Villejuif, France ;,* ***2.*** *CIRAD, UMR AGAP Institut, F-34398 Montpellier, France - UMR AGAP Institut, Univ Montpellier, CIRAD, INRAE, Institut Agro, F-34398 Montpellier, France,* ***3.*** *IFB-core, Institut Français de Bioinformatique (IFB), CNRS, INSERM, INRAE, CEA, 94800 Villejuif, France.,* ***4.*** *BGE - Laboratoire Biosciences et bioingénierie pour la santé, 17 avenue des Martyrs 38 054 Grenoble cedex 9 - France, CEA : DSV (Centre de Saclay Centre de Grenoble Centre de Cadarache etc - France),* ***5.*** *INRAE, BioinfOmics, Plant Bioinformatics Facility, Universite Paris-Saclay, Gif-sur-Yvette, Île-de-France, 78026, France*

Abstract

ELIXIR, the European research infrastructure for life science data, plays a central role in promoting open science by coordinating resources, standards, and thematic communities to make research data FAIR and facilitate their sharing across Europe and beyond. The French ELIXIR node, the Institut Français de Bioinformatique (IFB), actively contributes to these efforts by participating in the development of the Research Data Management (RDMkit), an initiative led by the ELIXIR Research Data Management community.

This poster provides an overview of RDMkit, an online toolkit that highlights best practices and supporting resources for research data management in accordance with the FAIR principles. RDMkit also serves as a comprehensive guide covering the entire data lifecycle. It is structured through dedicated pages organised by country, professional role, or scientific domain.

The poster will particularly focus on recently developed resources: the Data Stewardship Handbook, an operational manual for data stewards, and the Research Data Management Maturity Model (RDM Maturity Model), a self-assessment and improvement framework designed for research institutions.

These resources are open, collaborative, and continuously updated by the ELIXIR community, promoting reuse, training, and the open data valorisation, code, and software.

URL

https://rdmkit.elixir-europe.org/fr_resources

OpenMetaBar & BarCodeR: two complementary tools for metabarcoding analyses

Poster

Matéo Léger-Pigout [1], ***Sophia Marguerit*** [1], ***Sylvie Warot*** [1], ***Ionela-Madalina Viciriuc*** [1], ***Nicolas Ris*** [1], ***Etienne G.J. Danchin*** [1], ***Corinne Rancurel*** [2]

1. *Institut Sophia Agrobiotech (UMR1355), INRAE, Université Côte d'Azur, 400 route des chappes, 06903, Sophia Antipolis, France,* ***2.*** *PHYBAC (EMR7006), CNRS, INRAE, Université Côte d'Azur, 400 route des chappes, 06903, Sophia Antipolis, France*

Abstract

Metabarcoding is a powerful molecular approach for characterizing the taxonomic composition of biological communities, including bacteria, archaea, and eukaryotes. However, its implementation relies on a complex analytical workflow involving data preprocessing, variant inference, taxonomic assignment, statistical analysis, and visualization. In practice, this workflow often combines heterogeneous tools, multiple scripts, and software packages, increasing processing complexity, limiting traceability, weakening reproducibility, and extending analysis time. To address these challenges, we present two complementary tools currently under development, OpenMetaBar and BarCodeR, designed as components of a unified software ecosystem that structures, standardizes, and strengthens the entire metabarcoding analytical continuum.

OpenMetaBar is a reproducible pipeline implemented in Nextflow, designed to automate processing from raw sequence files to standardized phyloseq objects. Based on a design file describing samples, FASTQ files, barcodes, primers, and metadata, it orchestrates input parsing, demultiplexing, sequence filtering, and the automatic generation of intermediate files required for downstream analyses. Depending on the data type and marker, it then runs adapted analytical branches up to taxonomic assignment and produces directly usable phyloseq objects.

BarCodeR is an interactive application developed in R Shiny, specifically designed to exploit directly the phyloseq objects generated by OpenMetaBar. Dedicated to data loading, editing, filtering, statistical analysis, and visualization, it extends the pipeline workflow without disruption. Easy to use yet highly configurable, it supports the manipulation of one or several phyloseq objects simultaneously, integrates abundance tables, taxonomy, metadata, sequences, and phylogenetic trees, and provides access to the main statistical analyses used in metabarcoding. It also offers extensive graphical customization, an overview of generated figures, and a history and logging system ensuring full traceability and reproducibility.

Together, OpenMetaBar standardizes and produces analysis-ready objects, while BarCodeR directly exploits them within an interactive environment for exploration, statistical interpretation, and figure production.

Optimizing de novo assembly of RCA-enriched circular ssDNA viral genomes using long-read sequencing

Poster

Pakyendou Estel NAME [1], ***Ezechiel TIBIRI*** [2], ***Fidele Tiendrebeogo*** [3], ***Angela ENI*** [4], ***Justin S. PITA*** [4]

1. *Laboratory of Virology and Plant Biotechnology, Institute of Environment and Agricultural Research (INERA), 01BP476 Ouagadougou, Burkina Faso,* ***2.*** *Laboratory of Virology and Plant Biotechnology, Institute of Environment and Agricultural Research (INERA), 01BP476,* ***3.*** *Central and West African Virus Epidemiology (WAVE), Pôle scientifique et d'innovation de Bingerville, Université Félix Houphouët-Boigny (UFHB),* ***4.*** *Central and West African Virus Epidemiology (WAVE), Pôle scientifique et d'innovation de Bingerville, Université Félix Houphouët-Boigny (UFHB), Bingerville, Côte d'Ivoire*

Abstract

Rolling circle amplification (RCA) combined with Oxford Nanopore sequencing enables the reconstruction of circular single-stranded DNA (ssDNA) viral genomes from complex plant samples. However, RCA generates concatemeric molecules that complicate *de novo* assembly, often resulting in artificial duplications, miscircularisation or structurally ambiguous contigs.

To address these challenges, we developed a Snakemake-based workflow[1] specifically engineered to resolve these artifacts and improve genome recovery through a multi-tiered assembly strategy (Figure 1). The pipeline implements a hierarchical assembly approach where Flye v2.9.3 [1] is utilized as the primary assembler, followed by Raven v1.8.3 [2] and Canu v2.3 [3] as complementary assembly approaches for samples unresolved after primary assembly. Contigs obtained were then polished in two steps using Racon v1.5.0 [4] for consensus correction, followed by Medaka v1.12.0[2] for neural-network-based base refinement, improving sequence accuracy and reducing typical Nanopore indel errors. To resolve concatemer-derived artifacts, TideHunter v1.5.5 [5] was applied to detect and split tandem repeats, enabling the recovery of monomeric circular genome sequences. Applied to 53 sweetpotato (*Ipomoea batatas*) field samples, the pipeline recovered 113 circular ssDNA viral genomes [6-7]. The primary Flye-TideHunter module yielded 10 complete and 43 partial Sweet potato leaf curl virus (SPLCV) genomes, 50 complete and 2 partial Sweet potato leaf curl deltasatellite (SPLCD) sequences, plus partial assemblies of Sweet potato symptomless virus 1 (SPSMV-1), Pepper yellow vein Mali virus (PepYVMV), Cotton leaf curl Gezira alphasatellite (CLCuGeA), and Cotton leaf curl Gezira betasatellite (CLCuGeB). Complementary assemblers recovered 18 additional complete SPLCV genomes, one complete PepYVMV genome, and 3 full-length CLCuGeA and CLCuGeB sequences.

This hierarchical workflow improves the reconstruction of RCA-enriched circular viral genomes from long-read sequencing data and provides a robust, reproducible framework for accurate and scalable characterization of complex circular viromes in plant-associated pathosystems.

URL

https://github.com/etibiri/denovo-assembly-pipeline

Optimizing Grapevine Fanleaf Virus Diagnostics: A Statistical Model for Representative Sampling in Infected Vineyards

Poster

Eva Chevalier [1], ***Pierre Mustin*** [2], ***Jean-Michel Hily*** [3], ***Wassim Rhalloussi*** [2], ***Carine Schmitt*** [1], ***Myriam Hagege*** [1], ***Isabelle Rachel Martin*** [4], ***Olivier Lemaire*** [2], ***Anne Sicard*** [2], ***Loup Rimbaud*** [5], ***Emmanuelle Vigne*** [2], ***Sélim Ben Chéhida*** [2]

1. *INRAE, Université de Strasbourg, UMR-A 1131 Santé de la Vigne et Qualité du Vin, 68000 Colmar, France,* ***2.*** *INRAE, Université de Strasbourg, UMR-A 1131 Santé de la Vigne et Qualité du Vin,* ***3.*** *Institut Français de la Vigne et du Vin,* ***4.*** *Institut Français de la Vigne et du Vin, 30240 Le Grau-Du-Roi, France, et Laboratoire Partenarial Associé Vitivirobiome, 68000 Colmar, France,* ***5.*** *INRAE, UR0407 Pathologie Végétale*

Abstract

French viticulture faces a long-standing and devastating challenge: infectious degeneration, caused primarily by grapevine fanleaf virus (GFLV). This viral pathogen is specifically disseminated by the ectoparasitic nematode *Xiphinema index*. Its segmented genome comprises two RNA (RNA1 and RNA2). Given the persistence of the virus in the soil and the limitations of current prophylactic measures, a deeper understanding of its genetic structure is essential to develop innovative control strategies. In this study, we assessed the genetic diversity of GFLV of 342 vines in six impacted plots across two major viticultural regions: Burgundy and Champagne. Using RNA-seq data from one quarter of the grapevines in each plot, and pooling two vines per sample, we identified a total of 389 RNA1 consensus sequences and 351 RNA2 consensus sequences. These sequences were classified into genotypes based on a minimum of 95% nucleotide sequence identity. We were then able to characterize the rarefaction and extrapolation of the GFLV genetic diversity for each plot. Given this dataset, we are developing a statistical model designed to optimize sampling protocols, specifically to determine the number of samples required to accurately capture GFLV genetic diversity, given the plot surface area and vine density. Finally, we are exploring the spatial structure of these genotype distributions to identify significant epidemiological patterns. These findings will provide a quantitative framework to characterize GFLV genetic diversity and spatial structuring within vineyard plots, improving the design of sampling strategies for epidemiological studies.

OSPIL, save your data, save the world

Poster

Loik Galtier [1], ***François Sabot*** [1], ***Daniel Salas*** [1]

1. *IRD montpellier*

Abstract

The availability of a data decreases by 17% per year due to the lack of software to read it. 80% of scientific publications would be lost in 20 years, and this does not take into account the effects of unusable data. This decrease in data prevents the reuse of more associated data and the program still need to be reproduced. This has a negative impact on time, money, and the environment. A simple internet research produces 0.02g of CO2, where one request of a known AI produces 4.92g. With the absence of metadata, the research is a lot more pricey. Open science is a solution to reduce these costs and allow for better use of the data. It's supported by the FAIR principles and the CARE principles.

We propose a pipeline allowing the synchronization of data (code) between reference platforms (Gitlab, Dataverse, Software Heritage and Hal) with the corresponding metadata generated automatically to have better persistence.

OSPIL uses the CI/CD pipeline from GitLab to be minimal. Users only needs to add the .gitlab-ci.yml file to their project and add their own keys in it (in a safe environment) to synchronize between every platform. The Gitlab administrator can predefine some global metadata to help and upgrade the synchronization of all projects in his group.

On average, only 14.35 metadata are entered into a dataverse. Our pipeline enables the completion of at least 18 metadata per project without the need to go on Dataverse. While many metadata cannot be synchronized due to the difference between the platforms and the nature of some metadata, we believe that helping scientists and developers can create a better environment and a better data management. The code get a Digital Object Identifier (DOI) (by dataverse) and a SoftWare Hash IDentifier (SWHID) (by Software Heritage).

pan2met: predicting metabolic networks at the scale of microbial pangenomes

Poster

Samuel Ortion [1], ***Violette Da Cunha*** [1], ***David VALLENET*** [1]

1. *LABGeM, Génomique Métabolique, Genoscope, Institut François Jacob, CEA, CNRS, Université Évry, Université Paris-Saclay*

Abstract

Prokaryotes, bacteria and archaea, are diverse and ubiquitous organisms that play major roles in human health as well as on soil and ocean ecosystems. Advances in genome sequencing technologies have generated vast genomic datasets, with several thousand genomes now available for some species. To address large-scale comparative genomics, pangenome analysis tools have been developed, providing key insights into microbial diversity, evolution, and adaptation by distinguishing core genes (shared by all strains) from accessory genes (variable and often linked to phenotypic traits). However, genome-scale metabolic inference methods are generally applied to individual genomes and may therefore fail to capture the full genomic variability of a species.

We are developing a tool, pan2met, to predict the metabolism of an entire species using its pangenome as input rather than a single genome. The method relies on pangenome graphs generated by PPanGGOLiN, in which nodes represent gene families and edges represent gene neighborhoods. Each gene family is functionally annotated to predict enzymatic activities and infer the corresponding reactome. A rule-based algorithm is then applied to predict the presence or absence of metabolic pathways. The method currently supports MetaCyc as a reference for reactions and metabolic pathways. As output, the tool provides a list of predicted pathways and their completion both at the pangenome level and for individual genomes.

Further developments will focus on these main axes:

- Integrate multiple sources of metabolic pathways and enzymes (e.g. KEGG, MetaCyc, Rhea, UniProt, etc.)
- Use the information of gene neighborhood in the pangenome as a criterion to enhance enzymatic activity and pathway prediction.
- Provide a collection of pangenome-scale metabolic networks for all species available in the PanGBank resource.

URL

https://github.com/labgem/pan2met

PanExplorer2 : Explore multi-scale genetic markers derived from pangenome graphs for interactive comparative genomics and diversity analyses.

Poster

Bayram Boukhari [1], ***damien meyer*** [2], ***alvaro perez quintero*** [3], ***Ian Quibod*** [3], ***Sébastien Cunnac*** [3], ***Alexis Dereeper*** [4]

1. *postdam university,* ***2.*** *CIRAD, UMR ASTRE, F-97170 Petit-Bourg, Guadeloupe, France,* ***3.*** *PHIM, CIRAD, INRAE, IRD, SupAgro, Université de Montpellier, F-34398 Montpellier, France,* ***4.*** *IRD, PHIM, South Green platform*

Abstract

Comparative genomics increasingly relies on pangenome representations to capture the full spectrum of genetic diversity within species. However, most current tools analyze individual classes of variation independently (e.g. genes, single nucleotide polymorphisms, tandem repeats, or structural variants) thus fragmenting evolutionary signals across incompatible data structures. Graph-based pangenomes provide a unified framework that preserves both sequence and structural relationships, yet few integrated platforms exploit this representation to systematically derive multiple genetic markers for downstream evolutionary and population analyses. Here we present **PanExplorer2**, an interactive and reproducible platform that extracts complementary phylogenetic and genotyping signals directly from pangenome graphs. Starting from a PGGB-derived graph, PanExplorer2 generates gene presence/absence matrices, core-SNP genotypes, VNTR profiles and structural variation calls, enabling multi-scale evolutionary inference within a single interface. These markers are coupled with built-in population structure analyses (sNMF) and FST calculation, association testing (Scoary), functional enrichment, and interactive visualization of synteny and graph substructures. The platform integrates a Dockerized Snakemake workflow, a Galaxy wrapper for HPC execution of the workflow, and a modern web interface built with Plotly Dash for exploring data. By bridging graph-based genome representations with downstream comparative genomics, PanExplorer2 enables scalable, reproducible, and exploratory analyses across bacteria and eukaryotes. The web server is available at https://panexplorer.southgreen.fr/ and code source is freely available here https://github.com/SouthGreenPlatform/PanExplorer.

URL

https://panexplorer.southgreen.fr/

PanGBank: a Database of Pangenome Graphs for Comparative Microbial Genomics

Poster

Jean Mainguy [1], ***Téo Lemane*** [2], ***Claudine Médigue*** [2], ***Alexandra CALTEAU*** [1], ***David VALLENET*** [1]

1. *LABGeM, Génomique Métabolique, Genoscope, Institut François Jacob, CEA, CNRS, Université Évry, Université Paris-Saclay,*

2. *1. LABGeM, Génomique Métabolique, CEA, Genoscope, Institut François Jacob, Université d'Évry, Université Paris-Saclay, CNRS, FRANCE 2. Institut Français de Bioinformatique, IFB-core, UAR 3601, CNRS, Villejuif, France*

Abstract

Pangenome analysis provides crucial insights into microbial diversity, evolution, and adaptation. However, publicly available downloadable pangenome resources to support such studies are currently lacking. In this context, we present PanGBank, a comprehensive database compiling pangenome collections constructed with the PPanGGOLiN software suite (1).

PanGBank currently provides pangenomes for more than 4000 prokaryotic species represented by $\geq$15 genomes, built through a reproducible NextFlow pipeline. The resource includes a RESTful API for programmatic access and a user-friendly web interface (https://pangbank.genoscope.cns.fr) for intuitive exploration by non-developer researchers. As of today, PanGBank comprises two pangenome collections built from genomes (originating from RefSeq and GenBank) of the GTDB taxonomic database (release R10-RS220) (2) ensuring broad coverage of microbial diversity.

PanGBank provides a scalable and updatable platform for pangenome exploration across microbial clades. By offering multiple curated pangenome collections, it fills a critical gap in the field and paves the way for broader, collaborative, and data-driven microbial genomics research. As a use case, PanGBank and the new features introduced in PPanGGOLiN version 2 are being applied to explore antibiotic resistance in *Acinetobacter baumannii* from a pangenome perspective.

References

1. Gautreau G, Bazin A, Gachet M, Planel R, Burlot L, Dubois M, et al. PPanGGOLiN: Depicting microbial diversity via a partitioned pangenome graph. PLoS Comput Biol. 2020 Mar;16(3):e1007732.

2. Parks DH, Chuvochina M, Rinke C, Mussig AJ, Chaumeil PA, Hugenholtz P. GTDB: an ongoing census of bacterial and archaeal diversity through a phylogenetically consistent, rank normalized and complete genome-based taxonomy. Nucleic Acids Res. 2022 Jan 7;50(D1):D785–94.

URL

https://pangbank.genoscope.cns.fr/

ParasiTE: detection of chimeric gene-transposon transcripts in plants

Poster

Jérémy Berthelier [1]

1. *Université de Haute Alsace, Laboratoire Vigne, Biotechnologies et Environnement, UR 3991, F-68000 Colmar, France*

Abstract

Transposons (or transposable elements) can constitute up to ~80% of plant genomes. Some are inserted within intragenic regions and can act as regulatory elements for nearby genes or be co-transcribed with them, producing chimeric gene–transposon transcripts. We developed ParasiTE, a bioinformatics tool designed to identify gene–transposon transcripts and predict their impact on gene RNA isoforms. ParasiTE predicts transposon sequences associated with alternative transcription start and end sites, as well as alternative splicing events. ParasiTE was controlled using an *Arabidopsis thaliana* dataset. We found that 8% of *A. thaliana* genes produce gene–transposon transcripts, some of which are regulated by epigenetic mechanisms and participate in plant responses to abiotic and biotic stress. We began analyzing the production of gene–transposon transcripts in *Vitis vinifera*, an economically important crop increasingly affected by climate change–related stressors.

URL

https://www.nature.com/articles/s41467-023-38954-z
https://github.com/JBerthelier/ParasiTE

PASTEC: An Automatic Transposable Element Classification Tool

Poster

Mohamad Yassine [1], ***Johann Confais*** [2], ***Marienne Wan*** [2], ***BARDET Etienne*** [2], ***Hadi quesneville*** [2]

1. *Saclay Plant Sciences - Bioinfo,* ***2.*** *INRAE*

Abstract

Transposable elements (TEs) are major drivers of genome evolution and plasticity in eukaryotes. Their accurate classification is essential for understanding genome structure and for downstream genomic analyses. PASTEC (Hoede et al., 2014) is an automatic tool that classifies TE consensus sequences according to the Wicker hierarchical classification system.

PASTEC combines multiple lines of evidence to assign each consensus to a TE family: (i) similarity searches (BLASTn, tBLASTx, BLASTx) against reference TE databases , (ii) HMM profile from Pfam and RexDB matching for conserved protein domains (GAG, AP, RT, RH, Tase, etc.), and (iii) structural features such as terminal repeats (LTR, TIR), ORFs, polyA tails, and tandem repeats. A rule-based agent system aggregates these evidences with configurable weights to produce a classification and a confidence index (CI) for each consensus.

PASTEC is available both as a module within the REPET pipeline (TEdenovo) and as a standalone Snakemake pipeline. The standalone version offers flexible configuration, support for Singularity or Conda environments, and can process consensus sequences from any source. In this poster, we present PASTEC's classification workflow, its evidence-based decision rules, and the main improvements of the current version.

URL

1. Hoede C, Arnoux S, Moisset M, Chaumier T, Inizan O, Jamilloux V, et al. (2014) PASTEC: An Automatic Transposable Element Classification Tool. PLoS ONE 9(5): e91929. https://doi.org/10.1371/journal.pone.0091929
2. Wicker T et al. (2007) A unified classification system for eukaryotic transposable elements. *Nat Rev Genet* 8: 973–982. https://doi.org/10.1038/nrg2165

PasteurAIze: A Multi-Agent Platform for Secure Natural Language Biomedical Data Analysis

Poster

Zakary Azmani [1], Tom Perdereau [1], Charles-Maxime Doudy [1], Rémi Planel [1], Etienne Patin [2], Fabien Taieb [3], Amine Ghozlane [1]

1. *Institut Pasteur, Université Paris Cité, Bioinformatics and Biostatistics Hub, F-75015 Paris, France,* ***2.*** *Human Evolutionary Genetics, Institut Pasteur, CNRS UMR2000, Université Paris Cité, Paris, France,* ***3.*** *Institut Pasteur Medical Center, Université Paris Cité, Paris, France*

Abstract

Background & Objectives:

Access to biomedical knowledge and analysis of complex, multi-modal datasets across immunology, microbiome, and clinical research remain challenging for both domain experts and non-specialists. Traditional approaches demand specialized programming skills and significant time investment, creating barriers that limit participation of non-specialists. Analytical workflows must respect data protection standards (GDPR). We present **PasteurAIze**, an ongoing project developing a multi-agent AI platform that democratizes data access and significantly reduces analysis time through natural language interaction with institutional datasets, while maintaining data security and reproducibility standards.

Methods: PasteurAIze uses a **Main Agent** coordinating four agents via the Model Context Protocol (MCP):

1. **Data Agent:** Database schema descriptions, metadata enrichment, and SQL query generation
2. **Search Agent:** Literature research via RAG with Google Scholar integration
3. **Global Coding Agent:** Statistical analysis with code execution (R/Python) in secure sandboxes
4. **Plot Analysis Agent:** Automated visualization analysis with follow-up capabilities

Results:

We developed a working prototype and validated it on several Institut Pasteur research studies. Future work includes the **Milieu Intérieur cohort** and **EOP (Observatory Survey of Traveler Diseases) health surveillance study**. The platform successfully extracts data from study databases, executes statistical analyses in secure sandboxes, generates visualizations, and allows users to ask follow-up questions about specific plots.

Ongoing Work:

Current development priorities include: (1) MCP-based agentic workflow development, (2) Milieu Intérieur data analyses in systems immunology, (3) Pasteur Medical Center epidemiological surveillance (EOP project), and (4) project management and ethics clearance. Future phases will explore integration of standardized Agent Skills documentation frameworks.

Conclusion:

PasteurAIze provides a comprehensive platform for secure, reproducible AI-driven biomedical data analysis using open protocols and standardized practices, enabling natural language interaction with institutional datasets while maintaining rigorous data security standards.

PHAREOM: streamlining multi-omics for translational research

Poster

Olivier Feudjio [1], ***Virginie MOURNETAS*** [1], ***Emmanuel LABARONNE*** [1], ***Alexandra BOMANE*** [1], ***Marion CRESPO*** [1]

1. *ADLIN Science*

Abstract

The rapid proliferation of high-throughput sequencing and proteomics technologies has generated unprecedented volumes of biomedical data, yet their integration and interpretation remain major bottlenecks in translational research and precision medicine [1, 2, 3]. To address these challenges, we have developed PHAREOM, a scalable, reproducible platform designed to standardize multi-omics data processing and accelerate translational applications, including biomarker identification and target discovery.

PHAREOM serves a broad user base: biologists and clinicians can leverage their analytical capabilities without computational expertise, while bioinformaticians benefit from ready-to-run, standardized pipelines without the burden of managing computing infrastructure. This is enabled by a sovereign cloud environment fully operated by our team, ensuring data security, regulatory compliance, and on-demand scalability.

The platform provides a modular framework for large-scale omics analysis, built around three core principles: reproducibility, interoperability, and automation. PHAREOM leverages containerized workflows and workflow orchestration systems such as Nextflow [4], combined with community-validated nf-core pipelines [5], to process raw sequencing data from FASTQ files to quantitative matrices and quality control reports. Standardized metadata models and automated data ingestion ensure consistent dataset structure, enabling robust cross-cohort comparisons in accordance with the FAIR data principles [6].

Beyond data processing, PHAREOM integrates analytical modules for biological interpretation, including differential expression analysis [7], pathways and gene set scoring methods such as GSEA [8], and biomarker signature identification enabling robust comparison and validation of molecular signatures across datasets and cohorts [9].

In its current implementation, PHAREOM delivers end-to-end transcriptomic pipelines, from raw sequencing reads to actionable biological insights. Designed for extensibility, PHAREOM will progressively integrate genomic and proteomic data layers, establishing a unified multi-omics framework for translational research.

Pipeline for the detection and quantification of ribosomal RNA nucleotidic variants from long read Oxford nanopore sequencing datasets.

Poster

Allyson Moureaux [1], ***Baudouin seguineau De Préval*** [2], ***Michelle Scott*** [2], ***Virginie Marcel*** [1]

1. *Centre de Recherche en Cancérologie de Lyon,* ***2.*** *Michelle Scott Lab*

Abstract

Recent studies have shown heterogeneity in ribosome composition, including at the level of ribosomal RNAs (rRNA). Sequence variations from rDNA and rRNA have been described and indexed. An atlas of human rDNA and rRNA variants has been proposed based on long read sequencing in which the frequency of certain variants in rRNA was determined for different healthy/cancerous tissues. Available tools that enable easy detection and quantification of variants for rRNA remain scarce and studies on this subject are limited. Moreover, there is no universal gold standard pipeline that allows reproductibility for rRNA variants detection.

Thus, an automated rRNA variant detection and quantification pipeline will be developed for long-read seqencing issued from the Oxford Nanopore Technology (ONT). It will consist of three steps.

First, the pipeline will perform extraction of rRNA reads from total RNA sequencing by performing a non-stringent alignment using minimap2. Second, it will align rRNA long-reads to several rRNA variants found in the literature by modifying the RGA tool employing Needleman-Wunsch alignment to accommodate indels and gaps. Finally, it will include statistical quantification of variants.

This will enable detection of known and *de novo* variants.

To develop and test this tool, we have sequenced cancerous and healthy samples (ONT). In future, the pipeline can be used to determine the impact of rRNA nucleotide variations on the chemical modifications profile of rRNA in different cancer.

Indeed, rRNA also exhibits heterogeneity in its chemical modifications, including 2'O-ribose methylation and pseudouridylation. Although the function of chemical modification alteration is still in its infancy, it has been shown that they may contribute to the regulation of ribosome activity and various biological processes. However, the origins of rRNA chemical modifications in patho-physiological context remain to be deciphered. One hypothesis is that rRNA nucleotide variants impact snoRNA:rRNA pairing and thus chemical modifications of rRNA.

Polygenic architecture of morbid obesity in individuals of European ancestry : a UK Biobank study

Poster

Lucille Herbay [1], ***Céline Wu*** [1], ***Anthony Haidamous*** [1], ***Claire Nominé-Criqui*** [2], ***Laurent Brunaud*** [2], ***David Meyre*** [1], ***Sébastien Hergalant*** [1]

1. *INSERM U1256, NGERE lab, Université de Lorraine,* ***2.*** *Visceral surgery department, CHRU de nancy*

Abstract

Obesity is defined by the World Health Organization (WHO) as an abnormal or excessive accumulation of body fat that may impair health. According to the WHO, 16% of adults worldwide were living with obesity in 2022, a prevalence that has more than doubled since 1990.

While overall obesity rates appear to be stabilising in some high-income countries, recent data suggest a shift towards more severe forms. These extreme phenotypes are associated with higher cardiometabolic risk and increased healthcare burden, underscoring the need to better understand their biological determinants.

Body mass index (BMI) is the most widely used phenotype in populational studies of obesity. While highly heritable and measured at scale, BMI is an imperfect proxy for adiposity as it does not distinguish between fat and lean mass in normal weight individuals although it does correlate with fat mass in severe to extreme cases of obesity (class 3).

Genome-wide association studies (GWAS) have identified approximately 1,000 frequent polymorphisms associated with BMI in the general population. In contrast, only few loci have been robustly associated with morbid obesity (BMI $\geq$ 40 kg/m²) in European adult. Thus it remains unclear whether this reflects the upper tail of the BMI distribution or involves distinct genetic architectures.

To answer that question , we conduct a case–control GWAS of adult morbid obesity in the UK Biobank, including 7,461 class 3 cases and 132,264 normal-weight controls of European ancestry, using TOPMed-imputed genotypes within a generalized mixed-model framework. We construct genetic blocks with variants in high linkage disequilibrium with significant hits (p<5e-8) and overlap those to identify shared and exclusive genomic regions linked to morbid obesity. We will also explore the pleiotropic potential of extreme-obesity-associated genes and link their effects to the comorbidities observed with these phenotypes.

Prédiction d'expression différentielle à partir des variants génomiques

Poster

Elliot Butz [1], ***Laurent BRÉHÉLIN*** [1], ***Charles Lecellier*** [1], ***Kévin Yauy*** [1]

1. *CNRS*

Abstract

Jusqu'à présent, la quasi-totalité des approches de deep learning (DL) qui prédisent un signal d'expression à partir de la séquence ADN utilisent une approche mono-génome où un modèle est entraîné à partir de séquences issues d'un unique génome, chaque séquence étant associée à un signal mesuré sur le même génome. Si ces approches sont performantes pour prédire l'expression des différents gènes d'un génome, elles sont peu efficaces pour le problème orthogonal, qui consiste à prédire l'expression du même gène dans différents génomes. Pour ce problème, d'autres approches moins basées sur des modèles linéaires sont plus précises. Contrairement aux méthodes de DL entraînant un unique modèle, ces approches entraînent un modèle par gène, et prennent comme variables prédictives l'ensemble des variants de chaque génome. Bien qu'elles soient globalement meilleures que les méthodes de DL sur ce problème, ces méthodes ont le désavantage de ne prendre en compte que les variants présents dans les génomes d'entrainement.

Nous proposons une voie intermédiaire entre ces deux approches. Plutôt qu'entraîner le modèle sur un unique génome comme les approches de DL classiques, nous l'entraînons sur l'ensemble des génomes d'apprentissage en modifiant la problématique : au lieu de prédire l'expression du gène g d'un génome G, on s'appliquera à prédire le différentiel d'expression entre ce gène g dans le génome G et la moyenne d'expression du gène g dans l'ensemble des génomes d'apprentissage. Notre idée est d'utiliser les prédictions produites par un modèle de DL comme variables prédictives d'un modèle linéaire facilement interprétable. Ainsi, si un gène d'un génome est prédit différentiellement exprimé par le modèle linéaire, une analyse des variables prédictives suffira à identifier la ou les propriétés chromatiniennes qui induisent cette différence d'expression ; cela signifiera que des variants impactent ces propriétés, et que cela a un effet sur l'expression du gène étudié.

Preliminary evaluation of the Transfer Learning capabilities of MOTL for multi-omics cancer survival analysis

Poster

Arnaud Gloaguen [1], ***Vincent Le Goff*** [1], ***Edith Le Floch*** [1]

1. *Mathématiques et Statistiques, CNRGH, Institut de Biologie François-Jacob, CEA*

Abstract

The high-dimensionality of multi-omics datasets combined with the multiplicity of their interactions make their analysis challenging, especially in the field of prognosis. In an insightful study (Herrmann et al. 2021) where 13 survival analysis methods were compared on 18 cancer data-sets analyzed separately, they highlight the limit of molecular data for the improvement of a prognosis based only on clinical data. To tackle this limit, we added joint Dimension Reduction (jDR) methods in a follow-up work (Le Goff et al. 2025) and managed to show a statistical improvement, aggregated across all cancers, in favor of these methods against models based on clinical data only. Yet, the main factor driving prediction performances is the number of patients rather than the chosen methodology. Here, we will focus on a paradigm to virtually increase the number of samples, Transfer Learning (TL). In order to learn a general “cancer knowledge” that will be transferred to a targeted cancer and ease its survival analysis, the pre-trained model will be learnt on a pan-cancer multi-omics dataset. We start by evaluating TL in the context of jDR methods as they seem to perform best in this context, raising MOTL (Hirst et al. 2025), built on the jDR method MOFA (Argelaguet et al. 2018), as the only candidate. MOFA and MOTL were compared on the 7 smallest datasets analyzed by Herrmann et al. (2021), as they are the most likely to benefit from TL. Preliminary results need to be consolidated but mainly highlight one cancer in favor of MOTL and one against (only significant results after correction). Attempts to interpret these results enhance the need to determine a procedure to select the best subset of cancers out of which the pre-trained model will be learnt for the transfer to be optimal on a designated targeted cancer.

Preliminary work on the development of a Knowledge-Distillation based framework able to handle missing modalities in the context of multi-omics integration

Poster

Mary Savino [1], ***Alberto Bastero Anegon*** [2], ***Arnaud Gloaguen*** [3], ***Edith Le Floch*** [1]

1. *Mathématiques et Statistiques, CNRGH, Institut de Biologie François-Jacob, CEA,* ***2.*** *CentraleSupélec,* ***3.*** *Génomique Métabolique, Genoscope, Institut de Biologie François-Jacob, CEA - CNRS - Univ. Evry / Université Paris Saclay, 91000 Evry, France.*

Abstract

In oncology and other pathological fields, quantifying survival risk along with genomic biomarkers is essential to better characterize diseases. To do so, data can be collected from patients across different sources, ranging from DNA sequences to protein quantification, constituting what is known as omics data. However, not all patients have every type of omics data available. Inspired by a pre-existing methodology, we introduced an alternative approach using knowledge distillation to first train a large model using only complete datasets on the considered modalities, called the teacher. Then, distilling its knowledge through the training loss to many simple models called the students, taking as input data from all patients of one specific modality. The results are then aggregated in an integration module which is robust to missing modalities.

Thus, the overall framework, composed of the teacher, the students and the integration modules is capable of using incomplete datasets during both training and inference. This new method was designed using VAEs and an integration module (VCDN or concatenation) for the multi-omic fusion. It aims at solving a survival classification task, using as much of the available dataset as possible. For this purpose, the methodology was first applied on a breast cancer cohort (BRCA) of TCGA, to classify patients into two survival categories. This cohort contains incomplete datasets, which makes it particularly relevant for our study.

This poster presents the methodology and the first results of a comparison of different knowledge distillation (KD) approaches, illustrating their potential benefits.

Profiling the escape from X chromosome inactivation in endometriosis

Poster

Nur Syahirah Binte Ruhazat [1], ***Camille Berthelot*** [1]

1. *Institut Pasteur, Université Paris Cité, CNRS UMR 3525, INSERM U1351, Comparative Functional Genomics group, F-75015 Paris, France*

Abstract

Endometriosis is a chronic inflammatory gynaecological disease affecting approximately 10% of women worldwide. It is characterised by growth of endometrial-like cells outside of the uterus, forming ectopic lesions. The local microenvironment that sustains these lesions is defined, in part, by altered immune cells and reduced immune surveillance. X-chromosome inactivation (XCI) is the random transcriptional silencing of a single X chromosome in female embryogenesis. However, about 15% of XCI genes escape inactivation, including genes related to immune and inflammatory regulation. This defective XCI is implicated in auto-immune diseases and cancer. Given the central role of immune dysregulation in endometriosis and the link between XCI escape genes and auto-immune diseases, XCI may play a role in endometriosis pathogenesis, but this role has been so far unexplored. Using publicly available single-cell and single-nucleus RNA-seq data of endometrial tissue from individuals with endometriosis and healthy individuals, we aim to identify an endometriosis-specific XCI escape gene profile. We will analyse differentially expressed annotated XCI genes and then quantify allele-specific inactivated X chromosome expression of any escapee genes. We will compare these findings with single-cell RNA-seq data from endometriosis lesions for validation. Understanding the role of X-chromosome inactivation in endometriosis could help improve the understanding of the molecular mechanisms of disease.

Profylo: A Python Package for Phylogenetic Profile Comparison and Analysis

Poster

SCHOENSTEIN Martin [1], ***Yannis Nevers*** [1], ***Odile Lecompte*** [1]

1. *Université de Strasbourg*

Abstract

Phylogenetic profiling exploits patterns of presence-absence of orthologous genes across different species to study coevolution and predict functional links between genes. With the increasing number of available genomes, this approach is becoming more and more powerful, particularly for exploring gene functions in less characterized species. However, the lack of unified and/or accessible tools limits its use and the comparison of these methods.

In this context, we developed Profylo, a Python library dedicated to the automated analysis of phylogenetic profiles. Profylo combines metrics with different strategies from the literature to measure the similarity between profiles. We also implemented clustering methods to identify evolutionary modules, i.e. groups of co-evolving genes. Finally, Profylo also integrates functions for visually, functionally and statistically analyzing the obtained modules.

As a use case, Profylo was applied to the analysis of the human proteome. The library made it possible to rigorously benchmark the implemented similarity metrics, by measuring their ability to identify known functional links from human KEGG pathways, on different datasets. The most recent methods that take into account the phylogenetic relationships between species have proven to be more effective than traditional ones based on naive vector comparisons. We have also demonstrated the impact of the taxonomic composition of the datasets used on the performance of the various methods.

In addition, we identified evolutionary modules in the human proteome and found functional groups of genes, such as those involved in the CatSper complex, featuring remarkable evolutionary histories in eukaryotes.

These results demonstrate Profylo's capacity to efficiently analyse large-scale phylogenetic profiles, uncover gene relationships, and explore genotype–phenotype relationships. Future challenges in the field of orthology lie in adapting current methods to the massive increase in the number of available genomes, with the optimization of large-scale orthology inferences, including the development of deep learning-based approaches.

URL

https://doi.org/10.1007/s00239-025-10280-6

RDMkit efficiently manages metabarcoding and metagenomic data

Poster

Clara Emery [1]***, Hanna Koivula*** [2]***, Yvan Le Bras*** [3]***, Vincent Lefort*** [4]***, Lucas Leclère*** [5]***, Éric Pelletier*** [6]***, Erwan Corre*** [7]

1. *IFB-core, Institut Français de Bioinformatique (IFB), CNRS, INSERM, INRAE, CEA, 94800 Villejuif ; ABiMS bioinformatic platform, FR2424, CNRS/Sorbonne Université, Station Biologique de Roscoff (SBR), 29680 Roscoff,* ***2.*** *CSC - IT Center for Science Ltd., Life Science Center Keilaniemi, Keilaranta 14, Espoo, Finland,* ***3.*** *PNDB Data-Terra research infrastructure biodiversity data hub, MNHN UAR 2047 DoHNÉE, Concarneau marine station, Quai de la croix, BP 225, 29182, Concarneau, France,* ***4.*** *Univ. Grenoble Alpes, Univ. Savoie Mont Blanc, CNRS, LECA, FR-38000, Grenoble, France,* ***5.*** *Biologie Intégrative des Organismes Marins (BIOM), Sorbonne Université, CNRS, 66650, Banyuls-sur-Mer, France,* ***6.*** *Génomique Métabolique, Genoscope, Institut de Biologie François-Jacob, CEA - CNRS - Univ. Evry / Université Paris Saclay, 91000 Evry, France.,* ***7.*** *CNRS-Sorbonne University, Station Biologique de Roscoff, FR2424, ABiMS-IFB ; Sorbonne Université, CNRS, Integrative Biology of Marine Models (LBI2M), Station Biologique de Roscoff*

Abstract

The **RDMkit** (Research Data Management toolkit for Life Sciences) is an **ELIXIR**-driven international resource designed to help scientists manage their data according to best practices. It enables users to identify the best practices for research data management (RDM) in line with the **FAIR Principles**: *Findable, Accessible, Interoperable,* and *Reusable*. For **metabarcoding** data management, the lack of clear standards has been widely acknowledged in the literature, inducing initiatives to develop specific solutions (Shea et al. 2023; Takahashi et al. 2025).

During the development of the RDMkit Biodiversity Domain page, the ELIXIR Biodiversity community identified the need for a dedicated RDMkit Metabarcoding Tool Assembly page. Driven by the ELIXIR Biodiversity and Microbiome communities, the existing marine metagenomics Tool Assembly page was also expanded to additional biomes and updated to include cross-community tools.

Here we present the Metabarcoding and Metagenomics RDMkit Tool Assembly pages developed within the ELIXIR communities, that encompasses technologies and standards used in each community. This work was led by the Biodiversity and Microbiome communities but also benefited the Plant, Food & Nutrition, Research Data Management and the Interoperability Platform expertise.

Recovering informative multiplex contacts from chimeric Hi-C and Micro-C reads using a split-and-parse workflow

Poster

Samir BERTACHE [1], ***Laurent MODOLO*** [1], ***Daniel JOST*** [1]

1. *Laboratoire de Biologie et Modélisation de la Cellule, École Normale Supérieure de Lyon, CNRS, UMR 5239, Inserm, U1293, Université Claude Bernard Lyon 1*

Abstract

Chromosome Conformation Capture assays such as Hi-C and Micro-C generate a substantial fraction of chimeric reads that are only partially exploited by standard analysis pipelines. More specifically, ligation-derived read structures, including chimeric and multiplex configurations, complicate mapping and can result in the loss of informative contact signal. Here, we present a split-aware computational framework to recover and filter additional contacts from both Hi-C and Micro-C data.

Our approach uses assay-specific strategies. For Hi-C, reads are first split at identifiable ligation junctions, then mapped and filtered. For Micro-C, where ligation signatures are ot directly identifiable from the sequence, reads are first locally mapped, then split based on the structure of chimeric alignments, remapped, and finally filtered. In both cases, mapped fragments are reconciled into candidate multiplex structures, from which valid contact pairs are reconstructed and exported as '.pairs' files after assay-aware filtering. The implementation is designed for large datasets and combines multiprocessing with parallel compression/decompression backends to support scalable analyses.

Using simulated datasets, we show that this strategy yields a substantial increase in the number of recovered informative reads compared with standard processing. Beyond increasing pair yield, this framework also provides explicit access to detected multiplex structures for downstream analyses. Overall, this work suggests that chimeric and multiplex read structures represent an underexploited source of contact information in 3C-derived assays. The framework provides standard outputs together with explicit multiplex structures, enabling downstream quantitative analyses of higher-order contact organization.

URL

https://gitbio.ens-lyon.fr/LBMC/hub/parasplit
https://gitbio.ens-lyon.fr/LBMC/hub/microsplit

Refining a Knowledge Graph Embedding library for reproducibility: the example of KGATE

Poster

Célia Brahimi [1]***, Benjamin Loire*** [2]***, Anaïs Baudot*** [3]

1. *Aix Marseille University, INSERM, MMG UMR 1251,* ***2.*** *Neurology Therapeutic Area, R&D Servier Paris-Saclay Institut,* ***3.*** *CNRS*

Abstract

Knowledge graphs (KGs) are widely used to organize large heterogeneous relational datasets. Knowledge graph embedding (KGE) methods enable machine learning models to learn vector representations of entities and relations in order to perform prediction tasks. Numerous KGE models and software libraries have been developed to support these approaches. KGATE is a knowledge graph embedding library designed to integrate existing libraries—such as TorchKGE, PyTorch Geometric, and PyKEEN—and multiple KGE models within a unified framework.

During the development of KGATE, we identified several challenges arising from differences between existing implementations. In particular, libraries often rely on incompatible internal data structures, making their integration within a single framework difficult and requiring substantial standardization efforts. Moreover, implementations of the same models may differ across libraries or even diverge from the original articles describing them, with little or no documentation explaining these choices. For example, the TransE model relies on different dissimilarity functions depending on the library used. These discrepancies make it difficult to fairly compare model outputs or evaluate performance differences, even when identical hyperparameters are used. In addition, documentation is frequently incomplete, outdated, or missing, which limits usability, particularly for users with limited programming experience.

To address these issues, we redesigned the development process of KGATE by following recognized best practices for scientific software development. We standardized naming conventions, reviewed the entire codebase, and implemented comprehensive documentation using structured docstrings based on the numpydoc format. We also created an online documentation website using ReadTheDocs and provided tutorials and contributor guidelines. The public GitHub repository includes standardized contribution workflows and templates for issues and pull requests.

By promoting standardized implementations, clear documentation, and open development practices, KGATE aims to improve reproducibility, accessibility, and interoperability in the knowledge graph embedding ecosystem.

URL

https://github.com/BAUDOTlab/KGATE

Remarkable repeated sequences in one of the most compact vertebrate genome

Poster

Faustine Collignon [1], ***Hugues Roest Crollius*** [1]

1. *Equipe DYOGEN, Institut de Biologie de l'ENS (IBENS), Département de biologie, École normale supérieure, CNRS (UMR8197), INSERM (U1024), Université PSL, 75005 Paris, France*

Abstract

Genome sizes vary more than 100-fold in vertebrates, in a large part because of the density of transposable elements. Some species contain extremely low frequencies of TEs, like pufferfishes, which exhibit some of the smallest vertebrate genome. Repeats and transposable elements are rare in these compact genomes and are compartmentalized in specific heterochromatic regions. Interestingly, one of these pufferfishes, *Tetraodon nigroviridis*, has dozens of copies of a specific pseudogene in its genome and these copies are located in the same heterochromatic regions as minisatellite repeats and transposable elements. This pseudogene called Trapeze, is made up of the first exons of a gene called EZH2 and the last exons of a gene called TRAP-a, and shows a variable structure throughout the genome. A second pseudogene highly amplified in *Tetraodon nigroviridis*' genome is iSET, which curiously spans the last five exons of EZH2, and is also located in the same heterochromatic regions. Genomic mapping of iSET, Trapeze and transposable elements copies revealed their tight co-localization within heterochromatic regions, to the point that many TE-iSET-Trapeze copies appear linked in triplets, which are duplicated in tandem. We describe here the genome-wide distribution of these unusual structures, which suggest possible mechanisms for their amplification. Altogether, these analyses suggest that iSET, Trapeze and TE are part of a common tandem repeat unit whose amplification and structural diversification reflect the rearrangement dynamics of heterochromatic regions. This landscape is unique among pufferfishes, and the repeated nature of these pseudogenic sequences is in stark contrast to the paucity of transposable elements in the *Tetraodon nigroviridis* genome.

Reproducible SNP-based phylogenomics reveals the population structure of multidrug-resistant Salmonella enterica serovar Kentucky ST198 in Burkina Faso

Poster

Marguerite Edith Malatala NIKIEMA [1], ***María PARDOS DE LA GÁNDARA*** [2], ***Laetita FABRE*** [2], ***Véronique GUIBERT*** [2], ***Magali RAVEL*** [2], ***Estelle SERRE*** [2], ***Nicolas BARRO*** [3], ***Lassana SANGARE*** [4], ***François-Xavier WEILL*** [2]

1. *Laboratory of Virology and Plant Biotechnology, Institute of Environment and Agricultural Research (INERA), 01BP476,* ***2.*** *Centre National de Référence des Escherichia coli, Shigella et Salmonella, Unité des Bactéries Pathogènes Entériques, Institut Pasteur, France,* ***3.*** *Laboratoire d'Epidémiologie et de Surveillance des Bactéries et Virus transmissibles par les Aliments, Ecole Doctorale Sciences et Technologie (EDST), Université Joseph Ki-Zerbo, Burkina Faso,* ***4.*** *Unité de Formation et de Recherche en Sciences de la Santé (UFR/SDS)/ Ecole Doctorale Sciences et Santé (EDSS), Université Joseph Ki-Zerbo, Burkina Faso*

Abstract

The global dissemination of multidrug-resistant (MDR) *Salmonella enterica* serovar Kentucky ST198, driven by fluoroquinolone selective pressure, represents a major public health concern. In West Africa, the genomic diversity and evolutionary dynamics of this lineage remain poorly characterized.

Within an integrated One Health surveillance framework, 161 *Salmonella* isolates collected from clinical and food sources in Ouagadougou between 2017 and 2018 were whole-genome sequenced (WGS). Following initial screening, isolates belonging to serovar Kentucky were selected for detailed phylogenomic analysis and contextualized within the global ST198 population structure. A reproducible bioinformatics workflow (Figure 1) was implemented: read filtering using fqCleanER v21.10 (https://gitlab.pasteur.fr/GIPhy/fqCleanER), *de novo* assembly using SPAdes v3.6.0 [1], genotyping using cgMLST/HierCC in EnteroBase (http://enterobase.warwick.ac.uk/), and SNP calling through reference mapping against strain 98K using Snippy v4.6.0 (https://github.com/tseemann/snippy). Maximum likelihood phylogeny was reconstructed with RAxML v8.2.12 [2] (GTR+I+G, 1000 bootstraps). Antimicrobial resistance genes were identified using ResFinder v4.1 [3].

All Kentucky isolates belonged to the globally disseminated X1-ST198-SGI1 lineage and exhibited high-level ciprofloxacin resistance associated with mutations in *gyrA* (S83F, D87Y/G) and *parC* (S80I). SNP-based phylogenomics revealed intra-lineage structuring separating food-associated isolates (SGI1-K1) from clinical subclades carrying SGI1-P2 and SGI1-K4 variants. Two West African clusters suggest possible independent introductions followed by regional diversification [4].

This study highlights the value of reproducible SNP phylogenomics for investigating the evolution and spread of MDR bacterial clones. Sequencing data are available in ENA (PRJEB44192) and the workflow is shared on GitHub (https://github.com/EdithNi/Reproducible-SNP-based-phylogenomics) to support open genomic surveillance.

URL

https://github.com/EdithNi/Reproducible-SNP-based-phylogenomics

Retrieval-Augmented Generation over Genomic Reports in the ABRomics Platform: Towards AI-Assisted Antimicrobial Resistance Research

Poster

Thomas Mignon [1], ***Raphaël Tackx*** [1], ***Julie Lao*** [1], ***Amanda Dieuaide*** [1], ***Brieuc Quemeneur*** [2], ***Philippe Glaser*** [3], ***Claudine MEDIGUE*** [4], ***Alban Gaignard*** [2], ***Gildas Le Corguillé*** [5], ***Fabien Mareuil*** [6]

1. *IFB-core, Institut Français de Bioinformatique (IFB), CNRS, INSERM, INRAE, CEA, 94800 Villejuif, France,* ***2.*** *Nantes Université, CNRS, INSERM, l'Institut du Thorax,* ***3.*** *Institut Pasteur, Université Paris Cité, Unité EERA, 75015, Paris, France,* ***4.*** *CNRS UMR8030, Université Evry-Val-d'Essonne, CEA, Genoscope, LABGeM, 91000, Evry, France,* ***5.*** *ABiMS bioinformatic platform, FR2424, CNRS/Sorbonne Université, Station Biologique de Roscoff (SBR),* ***6.*** *Institut Pasteur, Université Paris Cité, Bioinformatics and Biostatistics Hub 75015 Paris, France*

Abstract

ABRomics is a web platform that analyzes antimicrobial resistance from bacterial genomes. It delivers reports containing resistance genes detection, assembly metrics, quality control (QC), plasmid typing, species identification and cgMLST typing. To help clinicians and researchers exploit these data, we present a Retrieval-Augmented Generation (RAG) system [1,2] that retrieves relevant documents, injects them into the prompt, and uses them to generate responses, composed of four components: (1) a triage agent routing natural language queries to specialized handlers; (2) a document RAG module operating over structured report chunks (narrative section, gene tables, QC); (3) an ontology RAG module using Antibiotic Resistance Ontology [3] as a knowledge source; and (4) a data agent translating natural languages questions into structured database queries (Text-To-SQL) for cross-sample statistics [4].

The system demonstrates three key capabilities: (i) routing and orchestration, where the triage agent dispatches queries to appropriate module (retrieving report-specific information, explaining technical quality metrics, or generating population-level visualizations); (ii) multi-modal response generation, returning text and interactive charts (gene frequency bar charts, geographic distributions), statistics summaries, and links to resources (ABRomics data, ontologies, platform documentation); and (iii) report enhancement, where the Large Language Model enriches raw genomic tables with context for researchers, public health officials, and clinicians.

This multi-module RAG system integrates automated genomic analysis, ontological knowledge, and interactive visualizations to provide clinicians and researchers with accessible insights into antimicrobial resistance data [5].

1. Gokdemir O, et al. HiPerRAG: High-Performance Retrieval Augmented Generation for Scientific Insights. PASC. ACM; 2025:1-13. doi:10.1145/3732775.3733586
2. He J, et al. Retrieval-Augmented Generation in Biomedicine: A Survey of Technologies, Datasets, and Clinical Applications. arXiv. 2025:2505.01146. doi:10.48550/arXiv.2505.01146
3. Hong Z, et al. Next-Generation Database Interfaces: A Survey of LLM-based Text-to-SQL. arXiv. 2024:2406.08426. doi:10.48550/arXiv.2406.08426
4. CARD-ARO. https://github.com/arpcard/aro
5. Giske CG, et al. GPT-4-based AI agents—the new expert system for detection of antimicrobial resistance mechanisms? J Clin. 2024;62:e00689-24. doi:10.1128/jcm.00689-24

URL

https://analysis.abromics.fr

Revisiting Effector Prediction Datasets using Protein Language Model Embedding Spaces

Poster

Eugeni Belda [1], ***Fanny Xie*** [1], ***Auguste Gardette*** [1], ***Jean-Daniel Zucker*** [1], ***Edi Prifti*** [1], ***Laura Gomez-Valero*** [2], ***Carmen Buchrieser*** [2]

1. *IRD, Sorbonne Université, UMMISCO, F-93143,* ***2.*** *Institut Pasteur, Unit of Biology of Intracellular Bacteria (BBI)*

Abstract

Legionella pneumophila, and more generally bacteria of the genus *Legionella*, are intracellular pathogens responsible for Legionellosis, a severe acute respiratory disease. Their pathogenicity relies on effector proteins secreted via a conserved Dot/Icm type IV secretion system to subvert host cell functions and establish intracellular replication. Identifying these effectors is a key challenge, as only around 300 have been experimentally validated out of an estimated repertoire exceeding 18,000 putative proteins across the *Legionella* genus.

While protein Language Models (pLMs) such as ESM-2 have revolutionized functional annotation by providing rich, high-dimensional embeddings, their success in binary classification tasks warrants closer scrutiny. In effector prediction, models leveraging these embeddings frequently reach near-perfect recall (e.g., >95–99\%). However, such exceptional performance often signals an 'easy' classification task due to inherent biases in the training data, rather than a breakthrough in biological signal detection. This raises a critical question: are pLMs learning the complex mechanisms of host-pathogen interactions, or are they simply exploiting the artificial statistical separability of existing datasets?

Here, we revisit these results through the lens of dataset composition. By projecting ESM-2 embeddings of the full L.*pneumophila* proteome into a low-dimensional space using t-sne algorithm, we confirm that the high performance of current models may partly reflect the artificial separability of training datasets (Legionella effectors vs. housekeeping proteins) rather than true biological discrimination. We propose an embedding-guided curation framework to construct more representative and challenging training sets, incorporating ortholog-derived candidates from evolutionarily related organisms as well as difficult non-effectors (sharing high embedding similarity with validated effectors identified from the literature), with the aim of providing more realistic benchmarks for effector prediction in bacterial pathogens.

RO-crate as a metadata source for the FAIDARE global federation

Poster

BARDET Etienne [1], ***Cyril Pommier*** [1], ***Celia Michotey*** [1], ***Raphaël Flores*** [1], ***Michael Alaux*** [1], ***Erik Kimmel*** [1], ***Maud Marty*** [1], ***Anne-Françoise Adam-blondon*** [1], ***Emma Leroypardonche*** [1]

1. *INRAE*

Abstract

Plant Research requires the discovery of heterogenous and dispersed datasets to prepare their integration and analysis. However, these datasets are often structured in different formats that are not strandardised and are stored in various databases across the world.

The FAIDARE data portal has been designed to tackle these issues. It enables dataset indexation using either a simplified and minimal metadata scheme, close to the Dublin core, or a more advanced one which conforms to the Breeding API standard. The latter maximise visibility whilst being in accordance with the FAIR principles(Findable, Accessible, Interoperable, Reusable).

FAIDARE allows the search of public data on plant biology from a federation of more than forty established data repositories over the world. It provides a unified one stop portal to search and access plant research data all over the world. It eases the findability and access of relevant datasets of most plant research data types, including genotyping, phenotyping and germplasms, ie genetic resources, via an easy-to-use web interface. It can also be accessed programmatically via web services conform to the Breeding API standard.

In parallel, the European research community gathered in the ELIXIR infrastructure for life science data to build the Research Object crate (RO-crate) standard. It allows the packaging of metadata, raw and derived data as well as the workflows and tools used for their analysis. It has been used in the European Variation Archive database hosted at EBI to expose all studies and Biosamples metadata. We will here present how the FAIDARE indexing tool suites now uses the EVA RO Crate datasource to enable genetic variation dataset discoverability in FAIDARE with Breeding API compliant metadata.

URL

https://urgi.versailles.inrae.fr/faidare/

Robust genotyping of grapevine fanleaf virus variants using amplicon-based Illumina sequencing

Poster

Pierre Mustin [1], ***Isabelle Rachel Martin*** [2], ***Wassim Rhalloussi*** [3], ***Shahinez Garcia*** [4], ***Myriam Hagege*** [3], ***Julie Kubina*** [5], ***Emmanuelle Vigne*** [3], ***Jean-Michel Hily*** [2]

1. *INRAE,* ***2.*** *Institut Français de la Vigne et du Vin,* ***3.*** *INRAE, Université de Strasbourg, UMR-A 1131 Santé de la Vigne et Qualité du Vin,* ***4.*** *Institut de Biologie Moléculaire des Plantes, Centre National de la Recherche Scientifique (CNRS), Université de Strasbourg, 12 rue du Général Zimmer, 67084 Strasbourg, France,* ***5.*** *Université de Strasbourg, 67000 Strasbourg, France*

Abstract

Although extremely informative, High Throughput Sequencing (HTS) RNA-seq may not be economically feasible for large-scale applications, limiting its use to explore plant protection strategies that require broad assessment of a pathogen's genetic diversity. This constraint is particularly relevant for cross-protection, in which infection of a plant by a primary virus prevents subsequent infection by another genetically related virus (superinfecting), thus requiring precise discrimination among viral variants [1]. To tackle this issue, we developed an Illumina-based amplicon sequencing approach (AmpSeq) to detect grapevine fanleaf virus isolates (*Nepovirus foliumflabelli*, GFLV), the causing agent of grapevine fanleaf degeneration and for which cross-protection research are currently investigated [1, 2]. Early HTS study on the GFLV-population composition [1, 3] enabled the design of cocktails of primers targeting a large diversity of GFLV (up to 27 clades representing an overall nucleotide diversity of $\pi \approx 0.10\text{-}0.12$) mainly found in French vineyards dedicated to cross-protection assays. Using these molecular tools, combined with Illumina libraries preparation and HTS sequencing protocols complemented with bioinformatics pipelines, we analysed and established the infectious status of greenhouse and field samples by confirming the presence of the expected primary-infecting GFLV variant(s) while identifying any additional GFLV sequences present in the sample. By fitting a binomial generalised linear model to the obtained detection rates, we were able to detect rare GFLV variants that accounted for less than 2% of the total reads obtained after a direct mapping (length fraction of 0.5 with similarity of 0.7) against all GFLV sequences from a sample. Moreover, detection sensitivity was not strongly influenced by either the RNA extraction approach or the enrichment method employed (detection threshold $\approx$ 0.49-3.21%), enhancing the cost-effectiveness and robustness of the method. Finally, beyond diagnostic, this approach can be used and adapted to characterise GFLV populations in vineyards and to demonstrate the effectiveness of cross-protection [4].

Scalable machine learning for large-scale genomic source attribution of L. monocytogenes

Poster

Isis Lorenzo [1], ***Zara Zulfiquar*** [1], ***Meryl Vila-Nova*** [1], ***Deborah Merda*** [1], ***Thomas Brauge*** [2], ***Benoit Durand*** [3], ***Sophie Roussel*** [4], ***Virginie Chesnais*** [1]

1. *SPAAD-ANSES,* ***2.*** *SANAQUA-ANSES,* ***3.*** *EPIMIM-ANSES,* ***4.*** *SEL-ANSES*

Abstract

Whole-genome sequencing has transformed microbial surveillance, producing large genomic datasets capturing pathogen diversity across time and space. Public repositories hosting several thousand of genomes pose challenges due to high-dimensional genomic feature spaces. *Listeria monocytogenes* (*Lm*), a major foodborne pathogen, requires source attribution to link clinical cases to food origins. Machine learning (ML) efficiently captures complex patterns for source attribution, though previous studies relied on small, geographically and temporally limited datasets.

Here, we evaluated ML-based source attribution models using a standardized dataset of 5,366 *Lm* genomes from the NCBI Pathogen Detection database (1998-2024, five continents) covering six food categories. Genomic feature matrices of increasing dimensionality (*d*) (e.g., cgMLST (*d*=1,748), pangenome genes (*d*=5,888), SNPs (*d*=183,083), and 21-mer (*d*=499,717)) served as input for Random Forest (RF) and Light Gradient Boosting Machine (LGBM) classifiers. Models were trained using an 80/20 stratified split, with hyperparameter tuning and 10-fold cross-validation. Model validation was further assessed on an independent dataset of 294 human-derived isolates from listeriosis outbreaks.

Source attribution performance ranged from F1=0.64-0.74, with LGBM achieving 8% higher F1 on 21-mers features, suggesting a greater ability to exploit complex genomic signals. Combining features slightly increased classification accuracy, indicating complementary predictive information across features. Recurrent errors between mammals' and birds' meat isolates suggested shared production environments, and aggregating related food categories improved performance up to 7%. Filtering low-confidence predictions using probability thresholds increased overall precision (F1=0.87), highlighting a "grey-zone" of uncertain predictions. Independent validation correctly predicted 72% of human-derived isolates, demonstrating ML relevance for public health. Nevertheless, we found a decrease in prediction accuracy with temporal and geographical distance using both test and validation datasets. Overall, this study demonstrates the potential of ML to leverage genomic Big Data for large-scale source attribution, guiding food safety investigations and emphasizing the need for continuous retraining with diverse, up-to-date genomic data.

URL

Scientific Workflow Reuse in Practice: An Empirical Study of Nextflow Pipelines

Poster

Lénora Buggenhoudt [1], ***George Marchment*** [2], ***Frédéric Lemoine*** [3], ***Sarah Cohen-Boulakia*** [1]

1. *LISN, Université Paris-Saclay,* ***2.*** *Université Paris-Saclay, CNRS, Laboratoire Interdisciplinaire des Sciences du Numérique,* ***3.*** *Institut Pasteur*

Abstract

The growing scale and complexity of biological data analyses have led to the large adoption of scientific workflows, which structure analyses as pipelines of steps to improve automation, reproducibility, and transparency. Workflows are also expected to support reuse, either by reusing entire pipelines or individual workflow steps. However, how workflows are reused in practice remains poorly understood.

Fifteen years ago, an empirical study of 898 Taverna workflows from myExperiment analyzed workflow and processor reuse and reported limited reuse and a strong concentration of reuse and authorship among a small group of developers. More recently, a second study analyzing Nextflow and Snakemake workflows suggested that curated repositories such as nf-core could foster better reuse.

Motivated by these developments, we revisit workflow reuse by replicating and extending these two studies on a large dataset of 3,021 Nextflow workflows collected from GitHub. Using automated workflow analysis tools, we examine reuse at the workflow and process levels, as well as the role of contributors and community-driven workflows.

Our results highlight important evolutions in workflow reuse practices, including changes in reuse patterns, contributor dynamics, and the growing influence of community-driven workflow ecosystems.

URL

https://gitlab.liris.cnrs.fr/sharefair/jobim2026-workflow-resue - the appendice of the paper provides all the additionnal information

scRAW: Representation learning for rare cell population identification

Poster

Fabien Bidet [1], ***Victoria Bourgeais*** [1], ***Loann Giovannangeli*** [1], ***Patricia Thébault*** [1]

1. *Univ. Bordeaux, CNRS, Bordeaux INP, LaBRI, UMR 5800*

Abstract

Single-cell RNA sequencing (scRNA-seq) data are essential for describing tissue heterogeneity, but rare cell populations remain difficult to identify. This challenge affects learning-based clustering methods because single-cell data are high-dimensional, sparse, noisy, imbalanced, and often affected by batch structure. In deep learning models, optimization is largely driven by abundant cell types, such that rare cells often contribute too weakly to organize the latent space. To address this, we present scRAW, a rare-population autoencoder framework for scRNA-seq clustering. scRAW combines standard preprocessing with a two-phase deep learning-based training strategy. In the first phase, all cells contribute equally to the reconstruction objective. In the second phase, cell-specific weights are computed from the latent pseudo-cluster size and local density such that small cell groups contribute more strongly to the training. A triplet-based regularization then increases the saliency of high-weight cells in the latent space, while adversarial batch correction can be added when batch labels are available. The learned representation is finally clustered with a state-of-the-art algorithm. We evaluated our approach on a human pancreas benchmark, where scRAW reached mean ARI = 0.9005± 0.0314 and mean ACC= 0.9087± 0.0215 across 15 independent evaluations. scRAW also ranks first on several label-agreement, rare-population, internal-structure, and batch-effect metrics. These results support custom rare-cell reweighting as an effective framework-level strategy that improves rare-cell identification without degrading overall clustering quality.

URL

https://github.com/Fabienbdt/scRAW

SnakeVir: A Snakemake Workflow for Viral Metagenomics

Poster

Florian CHARRIAT [1], ***Antoni Exbrayat*** [1], ***Serafin Gutierrez*** [1]

1. *ASTRE, CIRAD*

Abstract

Viral metagenomics requires optimized analytical workflows to identify and classify viral sequences from complex environmental samples. Here, we present a step-by-step approach to enhance viral sequence detection and classification.

Here we present SnakeVir, a Snakemake-based workflow, designed for the analysis of viral metagenomic short-read data. It offers flexibility by allowing users to customize parameters and databases via the config.yaml file to adapt the workflow to specific datasets.

The pipeline begins with read processing, where sequencing adapters and low-quality reads are removed using Cutadapt. Reads matching ribosomal RNA sequences from Diptera and bacteria are filtered out through mapping with BWA. Following this, a co-assembly strategy is employed, combining data from multiple libraries to generate optimal contigs using Megahit et Cap3. Then viral sequences are identified using Diamond by comparing contigs to the NCBI nr database. To improve specificity, the workflow includes a filtering step to remove potential endogenous viral elements (EVEs) based on BLASTn searches against NCBI nt. Once viral sequences are identified, a quantification step is performed by mapping reads back to viral contigs to estimate their abundance in each sample

To refine taxonomic classification, SnakeVir applies a clustering approach using a label propagation algorithm, grouping contigs based on sequence similarity and abundance patterns across samples.

Finally, the workflow generates a comprehensive HTML report to facilitate data visualization and ensure reproducibility.

SNPer, a web app for annotated variant mining

Poster

Frédérique Bitton [1], ***Jacques Lagnel*** [1], ***Mathilde Causse*** [1]

1. *INRAE*

Abstract

High-throughput resequencing and genotyping generate large variant datasets, but biologists often struggle to extract biologically relevant variants from custom gene lists, genomic intervals, or QTL confidence regions. Here we present SNPer, a web-based application that enables rapid, interactive querying of ANNOVAR-annotated VCF files from multi-accession plant populations.

SNPer supports three complementary search modes: (i) gene list queries with fuzzy matching against gene IDs, (ii) genomic interval searches and (iii) haplotype-based QTL analysis where users specify peak patterns (A/B/neutral) across accessions to identify compatible variants. Results display variant coordinates, overlapping genes and functions, ANNOVAR effect predictions impact levels, and genotype matrices, with dynamic filtering by effect type, variant class (SNP/indel), and minor allele frequency recalculated for user-selected accessions.

The application ingests standard ANNOVAR-annotated VCFs, GFF3 gene annotations, and accession metadata into a normalized PostgreSQL database with spatial/text indexes for efficient querying. In order to facilitate the deployment, SNPer is containerized using Docker (Django/Gunicorn + PostgreSQL + Nginx) and includes a pre-populated test database from a tomato resequencing panel (Causse et al. 2014).

SNPer fills a gap between genome browsers (visualization-focused) and command-line tools (non-interactive), providing experimental biologists with an intuitive web interface to prioritize functional candidate variants directly from QTL mapping results. Source code, documentation, Docker image, and test data are freely available under CeCILL license from the INRAE forge *(https://forge.inrae.fr/gafl/i2b/dev_snper/snper.git)*.

SNPs functional annotation tools using eQTL and meQTL data

Poster

Lucie Troubat [1], ***Haibo Huang*** [1], ***Anja Estermann*** [1], ***Christophe Linhard*** [1], ***Raphaël Vernet*** [1], ***Florence Demenais*** [1], ***Emmanuelle Bouzigon*** [1]

1. *Université Paris Cité, Inserm, HealthFex, group of Genomic Epidemiology of Multifactorial diseaseS, Paris, France*

Abstract

Large-scale genome-wide association studies (GWAS) aim to identify genetic variants that underlie multifactorial diseases. These studies generate association results between a given disease or trait and millions of single nucleotide polymorphisms (SNPs). The vast majority of SNPs detected by GWAS as significantly associated with disease or trait are located in non-coding regions of the genome. Functional annotation of these genetic variants is a key step for understanding their biological implications in the phenotype under study. In that context, we developed two tools: Searcheqtl and Searchmeqtl, which integrate GWAS results with results of tissue-specific gene expression quantitative trait loci (eQTL) studies and tissue-specific DNA methylation quantitative trait loci (meQTL) studies. Searcheqtl takes as input a list of SNPs (rs numbers). The first step is to search for all SNPs in linkage disequilibrium (LD) with each SNP of the input list using a pre-defined genetic correlation threshold (r^2) and a reference panel. The second step involves investigating which of the SNPs identified at step1 are present in each of six eQTL databases from different tissues (e.g. lung, blood, skin and spleen), and adding gene information using Ensembl and HGNC. The final step is to generate a summary file combining all information (query SNP, LD SNPs, r^2, LD, eqtl p-value, eqtl z-score, gene, tissue, database, the best eQTL for a given gene for each tissue and database etc.). SearchmeQTL operates using the same procedures as Searcheqtl, but uses four meQTL databases from blood tissues, as well as a database (Ensembl) that provides information on genes located upstream and downstream of the gene of interest. Both tools are constantly being improved (e.g. error management, addition of new databases). To demonstrate the effectiveness of our tools, we applied them to GWAS results of eosinophil cationic protein and eosinophil-derived neurotoxin levels in asthma-ascertained families.

Spatial and transcriptomic profiling reveal cell-specific mechanisms of epilepsy in Focal Cortical Dysplasia Type II

Poster

Franz Dervis [1], ***Emilia Puig Lombardi*** [2], ***Reyes Castano-Martin*** [1]

1. *Institut Imagine,* ***2.*** *Institut Imagine, Université Paris Cité, INSERM U1163*

Abstract

Intra-patient heterogeneity and the rarity of pathological cell types pose significant challenges for characterizing Malformations of Cortical Development (MCD), such as Tuberous Sclerosis Complex (TSC) and Focal Cortical Dysplasia (FCD). While multiplex spatial imaging has successfully identified “microtubers” and characterized the morphology of dysmorphic neurons (DNs), these pathological cells remain difficult to detect in standard single-cell and single-nucleus RNA sequencing workflows, where they frequently fail to form discrete transcriptomic clusters.

In this study, we propose an analytical framework that shifts the focus from cluster discovery to continuous cell-state modeling. Using high-plex imaging results as a phenotypic “anchors”, we reanalyzed in-house snRNA-seq data alongside large-scale public multi-omic datasets to define a robust, reproducible DN transcriptional state score. Our approach utilizes UCell-based modules scoring and kNN-smoothing to capture a signal of DNs within excitatory neuron lineage, focusing on neurofilament-associated markers, mTOR pathway activation, and altered synaptic programs (GABA-chloride axis).

To ensure that these findings are not driven by sampling artifacts, we implemented a meta-analysis pipeline with reference mapping and label transfer across multiple platforms (10x 3′/5′ and Fixed RNA Profiling), enabling cross-dataset validation of the DN transcriptional state. Our results show that the DN state is significantly enriched in FCD type II compared to controls and is consistently associated with cytomegaly and stress response regulons (SCENIC). By harmonizing multiple cohorts, we provide a high-power molecular description of DNs that aligns with imaging-defined phenotypes. This work establishes a scalable bioinformatics strategy to bridge the gap between morphological “ground truth” and transcriptomic plasticity in neurological disorders.

Spatiotemporal mapping of cellular dynamics during epileptogenesis.

Poster

Raphaël Edery [1], ***Adrien Dufour*** [1], ***Baptiste Porte*** [1], ***Christophe Le Priol*** [1], ***Jeanette Nardelli*** [1], ***Guillaume Marcy*** [1], ***Cyril Degletagne*** [1], ***Andrée Delahaye-Duriez*** [1]

1. *INSERM*

Abstract

Epileptogenesis is the pathological process through which a healthy brain becomes capable of generating spontaneous seizures, involving profound molecular and cellular reorganization. In experimental models, this process unfolds in three phases: the acute phase (24 hours) characterized by neuroinflammation and neuronal death, the latent phase (7 days) a clinically silent period marked by intense molecular remodeling, and the chronic phase (56 days), during which spontaneous recurrent seizures emerge. Understanding how different brain regions and cell types contribute to this process remains a major challenge.

The objective of this work is to characterize the regional and cellular specificity of transcriptional responses during epileptogenesis by integrating two complementary single-cell datasets from pilocarpine-induced rat models. Specifically, we combine the multi-regional snRNA-seq atlas generated by Wang et al. with a snRNA-seq dataset produced by us.

The analysis pipeline begins with the annotation of Wang's dataset (hippocampus, cortex, and thalamus) using interspecies mapping with the MappMyCells tool (Allen Brain Institute). This approach assigns each nucleus to the Whole Brain Mouse taxonomy, providing a standardized framework for cell-type identification. An iterative cleaning process and refinement step are then performed through de novo reclustering to purify major cellular populations such as neurons, astrocytes and microglia.

Datasets are subsequently integrated using Robust PCA (RPCA) to reduce batch effects while preserving biological variability. Integration quality is quantitatively evaluated using LISI (batch mixing) and cLISI (conservation of biological identity) metrics, and optimal clustering resolutions are determined using Clustree.

The resulting integrated atlas enables the characterization of temporal and regional dynamics of cell populations. Finally, single-cell signatures will be projected onto Visium spatial transcriptomics data using CARD, providing a spatially resolved view of cellular remodeling and dynamic programs within epileptogenic circuits.

Spirochase: An easy-to-navigate portal to explore proteomes in the Spirochaetes Phylum

Poster

Elodie Chapeaublanc [1], ***Rachel Torchet*** [2], ***Samuel García Huete*** [3]

1. *Hub de Bioinformatique, Institut Pasteur,* ***2.*** *Institut Pasteur,* ***3.*** *Instituto Ramón y Cajal de Investigación Sanitaria*

Abstract

Spirochaetes, ancient bacteria found in diverse environments, are widely distributed and offer a unique opportunity to explore ecological transitions. Despite their high diversity and the presence of globally important pathogens such as syphilis (Treponema spp.), Lyme disease (Borrelia spp.), or leptospirosis (Leptospira spp.), Spirochaetes remain poorly characterized as a phylum. The Biology of Spirochetes Unit of Institut Pasteur Paris, examined a curated dataset of 172 spirochaetal genomes (all cultivable spirochete species) and conducted a comprehensive phylogenomic and functional analysis of Spirochaetes.

To facilitate the exploration of the results linked to this study (Huete et al.), a web application was developed to easily navigate the latest updated phylogeny tree, explore the annotations of the reference proteomes, or perform a simple BLAST search for a protein of interest.

On the technical side, the interface was designed by a UX/UI designer to facilitate user navigation, dynamic search, and interactive visualization. The implementation of the interface was done in R using the Shiny package. This approach allowed the integration of scientific code and web interface using the same language. The web application displays an interactive Plotly plot, permitting a live exploration of the figure from the paper. The BLAST search function rapidly identifies proteins of interest, with results shown in a dynamic, selectable, and filterable Reactable table. After selecting an entry in the BLAST table, the distribution is highlighted in the phylogenetic tree. Results and data can be exported and downloaded.

In summary, Spirochase provides an accessible platform for exploring the proteomes of the Spirochaetes phylum, enhancing the understanding of these significant bacteria.

URL

URL 1 : https://doi.org/10.1101/2025.07.04.663154

URL 2 : https://spirochase.pasteur.cloud/

Stability selection algorithm for biomarkers selection in high dimensional data

Poster

Thomas Carvaillo [1], ***Romain Torres*** [1], ***Margot Zahm*** [1], ***Sébastien Déjean*** [2], ***Olivier Joffre*** [1]

1. *Toulouse Institute for Infectious and Inflammatory Diseases (Infinity), INSERM UMR1291 – CNRS UMR5051,* ***2.*** *Institut de Mathématiques de Toulouse, UMR5219, CNRS, UPS, Université de Toulouse*

Abstract

High-dimensional analysis, especially when the number of variables exceeds the number of individuals (the p >> n case), is common in high-throughput biology. Thus, we need, in routine analysis, to address overfitting, model instability and multicollinearity. To resolve these issues, we propose an approach based on Stability Selection. Our improvements focus on coefficient estimation and the reporting of measurement errors.This method combines the L1 penalization (LASSO model) with subsampling without replacement. For a set of regularization parameters that satisfy a user-defined error control (an upper bound called PFER), we compute, for each of them, the selection frequencies of each variable across the resampled dataset. Finally, the “stable variables” will be those with a selection frequency that is higher than a user-defined threshold for at least one regularization parameter. This approach focuses only on predictors that present a true relationship with the dependent variable, and avoids sample bias (with respect to the random choice of the training set).Once the variable selection is achieved, our proposal is to set the regularization parameter in a retroactive way to report measurements, by considering the highest that selects at least the “stable variables” with a probability greater than the threshold. Then, one can obtain an Out-Of-Bag MAE, RMSE, Adjusted R Squared, … measured on the unused samples across the different subsamplings.We apply this methodology to the identification of age-associated biomarkers, extracted from a panel of 824 biomarkers measured in 500 individuals from the INSPIRE-T cohort. Thus, we aim to propose a pipeline to analyze high-dimensional flow cytometry data to assess the relationship between biological markers and a continuous dependent variable. Above all, we seek a methodology that provides a parsimonious framework, which is more interpretable than purely performance-driven.

Statistical learning for predicting gene expression from transcription factor expression

Poster

Manal BEZIA [1], ***Etienne Delannoy*** [1], ***Marie-Laure Martin*** [2]

1. *IPS2,* ***2.*** *IPS2; MIA Paris-Saclay*

Abstract

Understanding how plants respond to combined environmental stresses is a major challenge in plant biology, particularly in the context of climate change. In this project, we analyze a dataset of 22 biological replicates of Arabidopsis thaliana, grown under four environmental conditions defined by two levels of atmospheric CO_2 and two temperature scenarios. Several molecular levels were measured in these plants: transcriptome, proteome, metabolome, and photosynthetic activity. Preliminary analyses showed that approximately 10% of genes exhibit an interaction effect between CO_2 levels and temperature on their expression, whereas this effect was not detected at the protein or metabolite level.

The project is to develop predictive models capable of estimating gene expression under combined stress conditions from the gene expression measured when only one environmental factor varies. This problem corresponds to a prediction framework based on partial data, which raises methodological challenges, particularly due to the high dimensionality of molecular data compared to the limited number of samples (p>>n). To tackle this problem, we reformulate the initial objective by exploring statistical learning methods for predicting gene expression from omics data.

Before predicting the expression in a condition which is not observed, we sought to predict the expression of a gene from the expression of the transcription factors. We also evaluated the ability to predict the response of a gene to a single stress from the response of the transcription factors, using generalized additive models as well as other approaches, including XGboost, and neural networks.

We identified the importance to pre-process the data to provide relevant results. We investigated the integration of biological knowledge about regulation. Finally, we implemented variable screening strategies to identify the most informative transcription factors for predicting the expression of a target gene.

Stress Adaptation Pathways and Druggable Vulnerabilities in MTUS1-Low Triple Negative Breast Cancer

Poster

Gwenn Guichaoua [1]***, Sylvie Rodrigues-Ferreira*** [2]***, Clara Nahmias*** [3]***, Véronique Stoven*** [4]

1. *Sorbonne Université, CNRS, IBPS, Computational, quantitative and synthetic biology, UMR 7238,* ***2.*** *Gustave Roussy Cancer Center - INSERM U981, Université Paris-Saclay, 94800 Villejuif, France - Inovarion, 75005 Paris, France,* ***3.*** *Gustave Roussy Cancer Center - INSERM U981, Université Paris-Saclay,* ***4.*** *Center for Computational Biology (CBIO), Mines Paris - PSL Research University - Institut Curie, PSL Research University - INSERM U1331*

Abstract

Triple-negative breast cancer (TNBC) is aggressive and heterogeneous, with limited targeted therapies. Low expression of the tumour suppressor *MTUS1* is associated with poor prognosis, yet the transcriptional programmes and therapeutic vulnerabilities linked to an *MTUS1*-low state remain insufficiently defined. It is thus crucial to identify robust *MTUS1*-associated tumour states that capture aggressive biology and nominate actionable vulnerabilities. Here, we describe a reproducible *MTUS1*-low state in TNBC by integrating multi-cohort tumour transcriptomics with functional genomics and pharmacogenomics to prioritise testable genetic dependencies and drug-response hypotheses.

We leveraged two complementary data types (patient tumours and TNBC cell-line models) and three analyses. (i) TNBC tumours from seven public RNA-seq cohorts were stratified within each cohort by *MTUS1* expression (lowest vs highest tertile). Hallmark pathway activity was quantified by GSEA, and reproducibility was assessed across cohorts and validation layers. (ii) In TNBC cell lines, functional liabilities were explored using DepMap Chronos CRISPR–Cas9 gene effect scores, testing gene-wise correlations between dependency and MTUS1 expression. (iii) In TNBC cell lines, pharmacogenomic associations were evaluated by correlating *MTUS1* expression with compound response in GDSC2 and PRISM, followed by target annotation and pathway-level enrichment analyses.

Our results are threefold. First, across cohorts, an *MTUS1*-low state is reproducible at the pathway level: MYC target programmes are consistently higher, together with signatures of proteostasis/UPR, DNA repair, and oxidative metabolism. Second, in TNBC cell lines, CRISPR dependency profiles point to functional reliance on the same modules, consistent with stress-buffering requirements in low-*MTUS1* contexts. Third, drug-response data (GDSC2 and PRISM) show exploratory signals of increased sensitivity to compounds linked to proteostasis, MYC-related transcriptional regulation, and replication-stress/DDR checkpoint inhibition (with targets enriched for cell-cycle/checkpoint control). Together, these results highlight testable therapeutic directions for *MTUS1*-low TNBC, notably strategies targeting MYC-associated transcriptional output, proteostasis capacity, and replication-stress/DDR checkpoint vulnerabilities.

URL

https://www.biorxiv.org/content/10.64898/2026.05.22.727134v1

Structuring and Interoperability of Thematic Data Management Plans for Research Entities

Poster

Sylvain Milanesi [1], ***Saliha Benzoudji-Beddek*** [1], ***Christophe Bruley*** [2], ***Jean-François Dufayard*** [3]

1. *IFB-core, Institut Français de Bioinformatique (IFB), CNRS, INSERM, INRAE, CEA, 94800 Villejuif, France.,* ***2.*** *BGE - Laboratoire Biosciences et bioingénierie pour la santé, 17 avenue des Martyrs 38 054 Grenoble cedex 9 - France, CEA : DSV (Centre de Saclay Centre de Grenoble Centre de Cadarache etc - France),* ***3.*** *CIRAD, UMR AGAP Institut, F-34398 Montpellier, France - UMR AGAP Institut, Univ Montpellier, CIRAD, INRAE, Institut Agro, F-34398 Montpellier, France*

Abstract

Since 2018, the French Institute of Bioinformatics (IFB) has been developing Data Management Plans (DMPs) tailored to research entities. This initiative, conducted in collaboration with the National Infrastructures for Biology and Health (INBS), and supported by the deployment of a dedicated instance of Data Stewardship Wizard by IFB (DSW@IFB) led to design an “Entity-DMP” model that allows information to be reused in project-level DMPs.

The choice of DSW was guided by its endorsement by the ELIXIR Interoperability Resources catalog; its active development (monthly updates); and its flexibility in creating and managing DMP templates. Moreover, IFB maintains and regularly improves a facility-DMP template structure.

At the national level, DMP-OPIDoR has been promoted by French research institutions to support mandatory DMP submissions for project proposals. Operated by INIST, this platform enables the creation of DMPs but does not allow users to modify the pre-defined structure, which limits customization.

Since 2025, the IFB has aimed to combine the strengths of both systems by leveraging the available APIs of DSW and DMP-OPIDoR. The long-term objective is to automatically extract information from an Entity-DMP in DSW to populate a project-DMP in OPIDoR. As a first step, the current effort focuses on transferring an Entity DMP directly from DSW to DMP-OPIDoR.

To achieve this, the team has developed a Python-based interface capable of extracting data, mapping corresponding fields, and formatting the output for seamless integration. The proposed approach transforms the DSW raw_questionnaire structure into a hierarchical JSON as a graph. Semantic annotations embedded in the DSW knowledge model are attached to the corresponding graph nodes and used, through its value,to provide a semantic bridge for automated mapping to the corresponding OPIDoR fields.

This work aligns with broader European initiatives such as OStrails (https://ostrails.eu/), which aims to define interoperability standards for research data management tools.

Study of mouse brain development transcriptome at transcript and exon level with ONT sequencing

Poster

William DESAINTJEAN [1], ***Martijn KERKHOFS*** [2], ***Julien COURCHET*** [2], ***Cyril Bourgeois*** [1]

1. *Laboratoire de Biologie et Modélisation de la Cellule, École Normale Supérieure de Lyon, CNRS, UMR 5239, Inserm, U1293, Université Claude Bernard Lyon 1,* ***2.*** *University Lyon 1, CNRS, INSERM, Physiopathologie et Génétique du Neurone et du Muscle, UMR5261, U1315, Institut NeuroMyoGène*

Abstract

Mouse brain development is a well-studied process that has been explored via numerous second-generation Illumina sequencing analyses to evaluate changes in gene expression and alternative splicing across developmental stages. However, third-generation sequencing technologies, such as that offered by Oxford Nanopore Technologies (ONT), which enable long reads to be produced, allow for a more insightful exploration of the transcriptome. This is achieved by building and quantifying the transcriptome at the transcript level and by showing the exact exon composition of each transcript, resulting from combinations of several splicing events. First, we used ONT sequencing to analyze the transcriptome of the wild-type (WT) mouse brain during postnatal development (days 1, 21, and 62). Bioinformatic analysis was performed using a customized version of the nf-core pipeline Nanoseq. The long reads were aligned using minimap2 and the transcriptome was built using two different tools, bambu and FLAIR, in order to compare their results. For each tool, the final transcript count matrix was analyzed using DESeq2 and DEXSeq. The gene expression was quantified using featureCounts and analyzed with DESeq2 to compare this results with the transcript analysis and previous results obtained with Illumina sequencing, at gene level, on the same subject.

In a second step, we sought to compare these data with ONT sequencing data from the brain of heterozygous (KO/WT) mice for the SON gene. SON encodes a splicing factor with an important impact on developing neurons, and it is mutated in the autism spectrum disorder-associated Zhu-Takenouchi-Tokita-Kim (ZTTK) neurodevelopmental syndrome. We will present the analysis pipeline and some of the results obtained from these analyses, which illustrate the numerous gene expression and alternative splicing variations that take place during this developmental process.

Summary of the MERIT/SFBI Survey on the Working Conditions of Bioinformaticians

Poster

Merit Bureau [1], ***Bureau SFBI*** [2], ***Erwan Corre*** [3]

1. *Réseau MetiER en bIoinformaTique MERIT,* ***2.*** *French Society of Bioinformatics SFBI,* ***3.*** *ABiMS*

Abstract

Eight years after the SFBI first raised the issue as part of the Profession Working Group[1], the SFBI and the MERIT Network's Profession Working Group proposed a survey on the working conditions of bioinformaticians to the community in February 2026.

The goal of this survey was to provide a comprehensive overview of the working conditions of bioinformaticians employed at various research institutes and private companies.

This survey, conducted among bioinformatics professionals, provides a detailed overview (250 responses) of the diversity of profiles, work environments, and key concerns regarding recognition, training, and salaries within the community in France.

Preliminary results indicate that while bioinformaticians are passionate about their scientific work and their autonomy, they nevertheless express a certain weariness regarding the precarious nature of their contracts, stagnant wages, and the lack of institutional recognition of their technical expertise (Figure 1).

The results of this analysis, presented in this poster, mark the first milestone in the joint SFBI/MERIT effort to elevate the status of the discipline at the national level.

URL

https://doi.org/10.5281/zenodo.5513972

Systematic and robust integration of bulk and single-cell RNA-seq to resolve the ion channel repertoire in Apis mellifera

Poster

Louis Closson [1], ***Matthieu Rousset*** [1], ***Thierry Cens*** [1], ***Pierre Charnet*** [1], ***Claudine Menard*** [1], ***Maxime Linard*** [1], ***Michel Boissac*** [1]

1. *IBMM UMR 5247 CNRS*

Abstract

Understanding the ion channel repertoire ("channelome") of *Apis mellifera* is a key challenge in both fundamental neurobiology and pollinator health research, as ion channels play central roles in neuronal function, behavior, and responses to environmental stressors. However, their characterization remains incomplete, particularly regarding cell-type-specific expression patterns, subunit composition, and interactions with auxiliary proteins.

Recent advances in single-cell RNA sequencing (scRNA-seq) provide unprecedented opportunities to investigate gene expression at cellular resolution in non-model organisms. Nevertheless, available honeybee scRNA-seq datasets remain relatively scarce and are often generated independently by different laboratories, under distinct biological conditions and using different genome annotations, limiting direct comparisons across studies.

To address these challenges, we developed a reproducible workflow for integrating publicly available honeybee brain scRNA-seq datasets comprising approximately 280,000 cells across multiple castes (queen, nurse, forager, and soldier), behavioral specializations, independent laboratories, and geographic origins. To ensure comparability across studies, gene identifiers were harmonized using a common genome annotation reference and standardized count matrices were generated.

Single-cell datasets were processed using Seurat, including quality control, normalization, highly variable gene selection, dimensionality reduction (PCA), batch-aware integration using Harmony, graph-based clustering (SNN-Louvain), and UMAP visualization. Particular attention was given to metadata definition and batch correction to account for both technical and biological variability across datasets.

To assess the robustness of the workflow, we systematically compared inferred cellular populations and channelome expression profiles across independent studies. Cross-study validation revealed reproducible cellular populations and conserved expression patterns despite substantial dataset heterogeneity.

Using the integrated single-cell dataset, we characterized ion channel expression across diverse brain cell populations, identified conserved channel co-expression patterns, and generated candidate hypotheses for channel assemblies. These results provide a valuable resource for future studies of neuronal signaling, channel diversity, and insecticide targets in *Apis mellifera*.

T2T genome assembly of a basal nematode reveals a complete epigenetic regulation toolkit and a horizontal gene transfer from plant

Poster

Julia Truch [1], karine Robbe-Sermesant [1], Arthur Péré [1], Dominique Colinet [1], Corinne Rancurel [2], Elodie Drula [3], Martine Da Rocha [4], Laetitia Perfus-Barbeoch [1], Erçan Seckin [4], Ulysse Julien-portier [1], Celine Lopez-roques [5], Carole Iampietro [5], Amalia Sayeh [5], Marie Gislard [5], Roxane Boyer [5], Leo Luvisutto [5], Daniel Esmenjaud [4], Etienne G.J. Danchin [4], Cyril Van Ghelder [4]

1. *Institut Sophia Agrobiotech, INRAE, Université Côte d'Azur, CNRS, 400 routes des Chappes, 06903, Sophia-Antipolis, France,* ***2.*** *PHYBAC (EMR7006), CNRS, INRAE, Université Côte d'Azur, 400 route des chappes, 06903, Sophia Antipolis, France,* ***3.*** *INRAE, Aix Marseille Univ, BBF, Biodiversité et Biotechnologie Fongiques,AFMB,Architecture et Fonction des Macromolécules Biologiques, USC1408, Marseille, France,* ***4.*** *Institut Sophia Agrobiotech (UMR1355), INRAE, Université Côte d'Azur, 400 route des chappes, 06903, Sophia Antipolis, France,* ***5.*** *GeT-PlaGe, Genotoul, INRAE, US1426, Castanet-Tolosan, France*

Abstract

The first telomere-to-telomere genome assembly for the order Dorylaimida, an early-branching nematode lineage was obtained using HiFiasm with PacBio HiFi long reads combined with the Nanopore longest reads and Hi-C Illumina contact data. Then we used Juicer from 3D-DNA with the Hi-C data to scaffold the contigs using the contact map. For each assembly, final scaffolding was manually edited using JuiceBox with the help of nematode telomere motifs identified using TIDK. Three assemblies were obtained, one for each of the randomly separated haplotype and a merged assembly. The merged assembly comprised the 10 expected chromosomes with telomeres at both ends. A substantial portion of haplotype-specific k-mer as detected by Kat spectrum can be explained by the heterozygosity of ~0.9% as estimated by Genoscope2.

Then phylogenetic analyses of the *X. index* predicted proteome were performed to search for horizontal gene transfers (HGT) because some were discovered previously in other parasitic nematodes. DIAMOND homology search results of the proteome were submitted to AvP to calculate the Alien Index (AI) and the Aggregate Hit Score (AHS) for each query sequence and semi-automated phylogenetic confirmation. 44 HGT were predicted: 36 originated from bacteria, 7 from fungi and one form plant. This constitutes the first documented case of a plant-derived HGT in a nematode genome.

X. index also harbors a complete set of canonical DNA methyltransferases, a full-length ATRX, DNA demethylase and CTCF. Using nanopore reads, we detected methylation at CpG genome-wide, which inversely correlates with gene expression and chromatin accessibility as in vertebrates. Among the 71 nematode proteomes analyzed using Interpro annotation and Blast analysis, the presence of this epigenetic toolkit seems specific to *X. index* clade of Dorylaimida. Together, these findings reveal a vertebrate-like epigenetic machinery in an early-branching nematode and call for a fundamental reassessment of the prevailing epigenetic paradigm in animals.

URL

https://doi.org/10.64898/2025.12.19.695367

Tackling the scRNA-seq integration challenge with a reproducible benchmarking framework

Poster

Sara Boughaba [1], ***Théo Noel*** [1], ***Maxime Mahé*** [1]

1. *The Enteric Nervous System in gut and brain disorders [U1235], INSERM, Nantes Université*

Abstract

Objective

Integrating single-cell RNA sequencing (scRNA-seq) datasets across experiments is essential for robust biological interpretation. Yet, integration methods differ widely in their ability to remove batch effects while preserving biological structure, and consistent comparisons remain difficult due to heterogeneous evaluation practices. We aimed to establish a systematic and reproducible framework to benchmark leading scRNA-seq integration methods.

Methods

We developed and implemented a modular evaluation pipeline built on the scIB (Single-Cell Integration Benchmark) framework using Snakemake for full reproducibility and scalability. The pipeline integrates eleven widely used methods, and evaluates them using thirteen complementary metrics assessing both batch effect removal and biological signal conservation, with subsets applied depending on method compatibility. To contextualize results, we included seven control baselines, ranging from raw unintegrated datasets to artificially optimal embeddings, defining empirical lower and upper performance bounds. The evaluation was conducted on two main datasets: the peripheral blood mononuclear cells (PBMC) test dataset and the human intestinal organoids (HIO) and HIOs with an enteric nervous system (HIOENS) dataset, used to assess the performance of the integration methods.

Results

Our large-scale benchmark reveals substantial variability across integration strategies depending on dataset characteristics and evaluation criteria. Some methods achieve strong batch mixing at the cost of biological fidelity, whereas others preserve biological structure but provide weaker batch correction. The inclusion of multiple baselines clarifies these trade-offs and highlights the theoretical limits of current approaches.

Conclusion

Our pipeline offers a transparent and extensible framework for quantitative comparison of scRNA-seq integration strategies. By standardizing evaluation across algorithms and datasets through defined metrics, a specific aggregation procedure, and rigorous baselines, it empowers researchers to make evidence-based method choices tailored to their experimental context.

URL

Currently in progress

Text2Meta: Automated Extraction and Structuring of RNA-seq Metadata from Scientific Publications using Large Language Models

Poster

Dylan Pin [1], ***Camille Rustenholz*** [1], ***Stéphanie Jaubert*** [2], ***Marco Moretto*** [3], ***Amandine Velt*** [1], ***Martine Da Rocha*** [2]

1. *1131 SVQV, INRAE,* ***2.*** *1355 ISA, INRAE,* ***3.*** *Fondazione Edmund Mach*

Abstract

Public omics data repositories such as NCBI SRA or EBI ENA contain millions of RNA-seq datasets, representing a valuable resource for large-scale comparative studies and meta-analyses. However, the reuse of these datasets is often severely limited by the lack of **structured, complete and standardized metadata** describing experimental conditions.

While sequencing data are systematically deposited in public archives, the associated biological context (genotype, tissue, treatment, developmental stage, stress conditions) is frequently incomplete in database records and instead dispersed across scientific publications and supplementary materials. This fragmentation makes large-scale integration and comparative analysis extremely challenging.

Here we present **Text2Meta**, a workflow designed to automatically extract and structure RNA-seq metadata from scientific articles and public repositories using recent advances in natural language processing and artificial intelligence. The pipeline combines several components: (i) metadata harvesting from ENA/SRA APIs, (ii) structured parsing of scientific publications and supplementary materials, (iii) **retrieval-augmented generation (RAG)** to identify relevant information within documents, and (iv) **large language models** to generate candidate metadata values supported by explicit textual evidence.

The approach is evaluated on a curated reference dataset composed of **more than 2,500 manually annotated RNA-seq samples of *Vitis vinifera*** derived from **86 publications**. This gold-standard dataset enables quantitative benchmarking of the extraction workflow and provides a validation framework to assess accuracy, coverage and robustness.

Text2Meta aims to produce standardized metadata tables with explicit provenance and traceability, allowing researchers to reliably link **RNA-seq runs to experimental conditions**. By facilitating large-scale metadata reconstruction, the method opens the door to meta-analyses of thousands of public transcriptomic datasets across biological contexts such as biotic and abiotic stresses, genotypes or developmental stages.

Beyond grapevine, the workflow will be transferable to species and omics datasets, improving FAIR data reuse and enabling **integrative analyses in plants** and other organisms.

The BioInformatics and Genomics (BIG) Platform at Institut Sophia Agrobiotech: Expertise and Resources for Multi-Omics Data Analysis in Plant Health Research

Poster

Martine Da Rocha [1], Arthur Péré [1], Matéo Léger-Pigout [1], Sophia Marguerit [1], Stephen Ambrogio [1], Etienne G.J. Danchin [1], Corinne Rancurel [2]

1. *Institut Sophia Agrobiotech (UMR1355), INRAE, Université Côte d'Azur, 400 route des chappes, 06903, Sophia Antipolis, France,* ***2.*** *PHYBAC (EMR7006), CNRS, INRAE, Université Côte d'Azur, 400 route des chappes, 06903, Sophia Antipolis, France*

Abstract

The BioInformatics and Genomics (BIG) platform of Institut Sophia Agrobiotech offers expertise in bioinformatics and solutions for processing, integrating, analyzing, and visualizing multi-omics data in the field of plant health and protection. The BIG platform is part of PlantBIOs (Biocontrol and Plant Biostimulation, Facilities and Expertise), labeled as a collective scientific infrastructure by INRAE. PlantBIOs offers equipment and expertise for studies ranging from the gene level to the whole agroecosystem scale with analytical tools (Imagery and Microscopy, Biochemistry and Mass Spectrometry, Video Phenotyping, Molecular Characterization, Bioinformatics and Genomics), experimental tools, and collections of rare biological resources.

Since the end of 2020, BIG has been an IFB contributing platform. The core team of the BIG platform is composed of three bioinformatics engineers: Martine Da Rocha, Arthur Péré, and Corinne Rancurel, the operational manager, as well as a systems administrator, Stephen Ambrogio. The core team is complemented by a scientific advisor, Etienne Danchin (senior scientist), and two fixed-term staff members, Matéo Léger-Pigout (PhD) and Sophia Marguerit (Engineer).

BIG's core expertise lies in comparative genomics, transcriptomics, small RNA analysis, and molecular evolution. More recently, the platform has expanded into epigenomics, pangenomics, and metabarcoding studies. The tools and resources developed by BIG are publicly available via its website, software forge, and integrative portals, and can be applied to similar challenges in other research areas. For example, the Alienness tool, which enables rapid detection of candidate horizontal gene transfers in genomes, has been used 2,125 times by 405 distinct users. The corresponding publications (Rancurel et al., 2017, doi:10.3390/genes8100248; Koutsovoulos et al., 2022, doi:10.1371/journal.pcbi.1010686) have been cited 71 times for Alienness and 32 times for AvP.

In addition to methodological developments, the platform offers support and training for biologists in the use of bioinformatics tools and pipelines developed by BIG.

URL

https://big.plantbios.sophia.inrae.fr

The Genotoul-Bioinfo platform

Poster

Christophe Klopp [1], ***Florent BLAISE*** [1], ***Philippe Bordron*** [2], ***Patrice Dehais*** [3], ***Vincent Dominguez*** [4], ***Nicolas ENJALBERT-COURRECH*** [1], ***Christine Gaspin*** [4], ***Maya GAWINOWSKI*** [1], ***Fabien GRAZIANI*** [1], ***Claire Hoede*** [2], ***Didier LABORIE*** [1], ***Théo MOSER*** [1], ***Philippe Ruiz*** [2], ***Martin RACOUPEAU*** [5], ***Marie-Stéphane TROTARD*** [1], ***Nathalie Vialaneix*** [5], ***Matthias Zytnicki*** [5]

1. *BioInfo Genotoul, MIAT UR875, INRAE, F-31326, Castanet Tolosan, France.*, ***2.*** *Université de Toulouse, INRAE, UR 875 MIAT, F-31320, Castanet-Tolosan, France,* ***3.*** *BioInfo Genotoul, GenPhySE, Université de Toulouse, INRAE, ENVT,* ***4.*** *Université de Toulouse, INRAE, UR 875 MIAT, F-31320,* ***5.*** *INRAE*

Abstract

The Genotoul-Bioinfo platform gathers a team of 17 engineers and researchers, with backgrounds in biology, computer science, and statistics. The platform offers 38 nodes with 128 cores and 2TB RAM, one node with 128 cores and 4TB RAM, >7PB storage, and 4 Nvidia A100 80GB GPUs. Our infrastructure also includes >950 installed software for bioinformatics uses, and ~150 data banks. We serve >1300 users, inside and outside INRAE and the Occitanie region, with an active support (>2000 tickets/year) to help them.

Our expertise includes (meta/pan)genome assembly and annotation, non-coding RNAs, and omics integration. We currently contribute to ~20 projects and we organize >10 days of training per year. We co-publish ~10 articles yearly, and we are mentioned in >80 articles each year.

These year, we will present three news. First, we acquired two new servers, with four L40S GPU on each server, equipped with 64 cores, and 1.5TB RAM. Second, we released a first version of PanAbyss (https://github.com/Pange31/PanAbyss). It models a pangenome graph in a Neo4J framework, and is able to query and visualize it interactively. A public instance is available (https://panabyss-dev.toulouse.inrae.fr). Last, we highlight a recent publication [1] published by members of the team. We built 25 high-quality metagenome-assembled genomes (MAGs), recovered from photogranules to treat synthetic wastewater (see Fig. 1). Cyanobacterial MAGs encoded photosynthesis and nitrogen fixation pathways, supporting internal oxygen and nitrogen cycling. Most heterotrophic MAGs contributed to nitrogen removal, highlighting the metabolic complementarity within photogranules studied for wastewater treatment.

1. Della-Negra O, Servien R, Milferstedt K, Hamelin J, Klopp C, Hoede C. Metagenome-assembled genomes from oxygenic photogranules obtained from photobioreactors treating synthetic wastewater. Environmental Microbiology 2026;15(3). Available from: https://doi.org/10.1128/mra.01310-25

URL

https://bioinfo.genotoul.fr/

The Virome@tlas project: from systematic harmonization of nucleotide sequence archives metadata to large-scale One Health applications

Poster

Elea Pauliat [1], ***Paul Tissot*** [1], ***Mélodie Fleury*** [1], ***Luca Nesterenko*** [1], ***Maël Rimeur*** [1], ***Stephane Delmotte*** [2], ***Romain Delunel*** [3], ***Julien DELLINGER*** [1], ***Vincent Lacroix*** [4], ***Caroline Leroux*** [5], ***Jérome Lejot*** [3], ***Romuald Marin*** [6], ***Dominique Guyot*** [1], ***Christine Oger*** [1], ***Matis Zouari*** [6], ***Christophe Blanchet*** [6], ***François Mialhe*** [3], ***Damien de Vienne*** [7], ***Hussein Anani*** [8], ***Laurence Josset*** [8], ***Jocelyn Turpin*** [5], ***Oldrich Navratil*** [3], ***Vincent Navratil*** [1]

1. *PRABI, Rhône-Alpes Bioinformatics Center, Université Lyon 1, 43 bd du 11 novembre 1918, Villeurbanne CEDEX 69622, France,* ***2.*** *Universite Claude Bernard Lyon 1, LBBE, UMR 5558, CNRS, VetAgro Sup, Villeurbanne 69622, France,* ***3.*** *CNRS 5600 EVS, Université Lumière Lyon 2,* ***4.*** *LBBE - Laboratoire de Biométrie et Biologie Evolutive - Université Lyon 1 - UMR 5558,* ***5.*** *IVPC UMR754, INRAE, Universite Claude Bernard Lyon 1, EPHE, PSL Research University, 69007, Lyon, France,* ***6.*** *IFB-core, Institut Français de Bioinformatique (IFB), CNRS, INSERM, INRAE, CEA, 94800 Villejuif, France.,* ***7.*** *Laboratoire de Biométrie et Biologie Évolutive, Lyon, France,* ***8.*** *Hospices civils de Lyon*

Abstract

Background: Leveraging **petabase-scale** nucleotide sequence archives offers a unique opportunity to characterize **viral biodiversity** and enhance "**One Health**" surveillance of emerging **pandemics**. However, such analyses require accurate metadata to identify **biases** that limit their interpretation. While international consortia advocate for standardized taxonomic, geographic, environmental, and temporal descriptions, integration remains challenging due to **inherent metadata incompleteness and heterogeneity**. Unlike manually curated datasets, which are often time-consuming and domain-specific, the Virome@tlas datalake architecture facilitates large-scale biogeographic analysis through an **automated** procedure built on **FAIR** principles, providing **annotated, harmonized** and **quality-scored metadata**.

Results: We developed a bioinformatic and geomatic pipeline to **annotate, correct, and assess the quality of host taxonomy, environmental, biological, and geographic metadata**. Processing over 50 million samples from the International Nucleotide Sequence Database Collaboration (INSDC) and China's Genomic Sequence Archive (GSA), we harmonized metadata following FAIR principles, such as the use of thematic ontologies or controlled vocabularies (e.g. ENVO, GOLD or UBERON). After post-processing, 71.4% of samples possess geographical coordinates with characterized precision, and 74% of host-associated metagenomes now have a valid scientific host name. Furthermore, 6.1M and 13.4M samples were assigned environmental (e.g., biome) and anatomical (e.g., gut) terms, respectively. The **Virome@tlas web platform** enables users to query, explore, filter, and visualize these **datasets** through **interactive geographical and taxonomic maps**.

Conclusion: Virome@tlas enables robust large-scale meta-analyses through **accessible, quality-informed metadata**, reaffirming the **importance of rigorous reporting** during submission. The consortium's future work will focus on i) **environmental re-contextualization** (*e.g.* climate, land-use, demography), ii) integrating **virus-host** and virus-sample databases (e.g., Vire, Serratus and NCBI STAT), and iii) metadata augmentation via **publication mining**. The annotated dataset will be available on our cloud-based platform in open-access (https://viromeatlas.univ-lyon1.fr/) until publication.

URL

https://viromeatlas.univ-lyon1.fr/

Transposable element dynamics in a conserved genomic segment of Pea revealed by comparative and pangenomic analysis

Poster

Mathieu Cartier [1], ***Jonathan Kreplak*** [2], ***Johann Confais*** [1], ***Judith Burstin*** [1]

1. *INRAE,* ***2.*** *Université Bourgogne Europe, Institut Agro Dijon, INRAE, Agroécologie*

Abstract

The genus Pisum has a particularly large genome (approximately 4.3 Gb), 83% of which is made up of repetitive elements. The genome size of wild and domesticated accessions varies from 3.9 to 4.8 Gb, suggesting a loss of transposable elements. To investigate this, a consortium recently generated chromosome-level assemblies of eight pea genomes, representing both wild and domesticated varieties, using Nanopore ultra-long reads and Hi-C. Transposable elements were annotated for each genome using DANTE, which recognises elements by their domains and structure. To better understand the impact of transposable element dynamics at the local level, panREPET was launched on regions spanning a few megabases, selected based on synteny.

A comparative genomics pipeline was developed to identify conserved genomic segments between two closely related species: Pisum sativum and Vicia faba. Briefly, 1:1 orthologous genes were used as an anchor for MCScanX results to estimate conservation between each syntenic segment. As a proof of concept, the most conserved segment was found in all eight available pea genomes using Liftoff. Gene order was conserved between different accessions, despite the presence of some recent putative tandem genes, which will be investigated. The total length of the segment varied among the accessions, ranging from 910,409 bp in domesticated and cultivated peas to 1,290,066 bp in wild accessions. The number of retrotransposable elements in the LTR family was similar across the eight accessions, but they represented a different proportion of the base sequence. For the wild accessions, 80% of the segments were LTRs, compared to 72% for the cultivated accessions.

A new version of panREPET, adapted to DANTE input, was used to analyse this segment's behaviour. The next step will be to reiterate the analysis on known loci of interest in order to explore the potential contribution of TE insertions to functionally relevant regions.

Understanding the Structural Landscape of Self-Incompatibility in Arabidopsis halleri through AI-Driven Protein Interaction Modeling.

Poster

Thomas Binet [1], ***Nessim Raouraoua*** [2], ***Althaf Saneen Karuvattil*** [2], ***Alice Namias*** [3], ***Julie Bouckaert*** [2], ***Marie Monniaux*** [3], ***Xavier Vekemans*** [3], ***Vincent Castric*** [3], ***Marc Lensink*** [4], ***Guillaume Brysbaert*** [2]

1. *University of Lille, CNRS, Inserm, CHU Lille, Institut Pasteur of Lille, US 41 – UAR 2014 – PLBS, 59000 Lille, France,* ***2.*** *University of Lille, CNRS UMR 8576-UGSF-Unité de Glycobiologie Structurale et Fonctionnelle, 59000 Lille, France,* ***3.*** *University of Lille, CNRS UMR 8198-Evo-Eco-Paleo, F-59000 Lille, France,* ***4.*** *Univ. Lille, CNRS UMR 8576-UGSF-Unité de Glycobiologie Structurale et Fonctionnelle, 59000 Lille, France*

Abstract

Self-incompatibility (SI) promotes genetic diversity in angiosperms by actively preventing self-fertilization. In Brassicaceae, like *Arabidopsis halleri*, this mechanism is governed by the highly polymorphic S-locus, which encodes the female S-receptor kinase (SRK) and the male pollen-specific ligand (SCR/SP11). The SI response is triggered by an allele-specific interaction between SRK and SCR, forming a tetrameric complex that dictates pollen rejection or acceptance (Ma et al. 2016).

While highly variable amino acid sequences—particularly in SCR—maintain different SI haplotypes, it remains unclear how such diverse sequences preserve the same functional interaction. Our previous phylogenetic studies (Chantreau et al., 2019) were constrained by a severe lack of experimental structural data (Berman et al. 2000). Furthermore, recent molecular dynamics observations suggesting partial compatibility between different haplotypes (Murase et al., 2020) relied heavily on limited homology models. Consequently, modern prediction tools (Akdel et al. 2022) remain underexplored for understanding these structural interactions across all haplotypes.

To bridge this knowledge gap, our project leverages recent AI advances in structural biology. Using AlphaFold-Multimer, AlphaFold 3, and our new tool, MassiveFold (Raouraoua et al., 2024)—which enhances AlphaFold's performance through massive sampling—we aim to predict 3D structures and protein-protein interactions. By building a comprehensive database of SRK and SCR variants and their complexes, we will map S-locus structures to identify the binding features controlling specific interactions. Ultimately, this AI-driven approach will elucidate the SI mechanism in *A. halleri* and provide deeper insights into receptor-ligand co-evolution.

URL

https://www.nature.com/articles/s43588-024-00714-4

Unlocking Microbial Dark Matter genomes using microscopy and Single-Cell Sequencing

Poster

Lucile Martin [1], ***Emilie Brivet*** [2], ***Caroline Monteil*** [2], ***Stéphanie Fouteau*** [1], ***Raphaël Méheust*** [1], ***Christopher Lefèvre*** [2]

1. *LABGeM, Génomique Métabolique, CEA, Genoscope, Institut François Jacob, CNRS, Université d'Évry, Université Paris-Saclay, 91057 Evry, 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*

Abstract

Patescibacteriota, a major component of the Microbial Dark Matter, are ultra-small bacteria with highly reduced genomes that are thought to depend on close associations with other microorganisms. They are predominantly non-cultivable and most of the Patescibacteriota genomes have been reconstructed from metagenomics studies thus losing the link between their symbiotic partners[1]. Magnetotactic bacteria are ubiquitous in aquatic environments and cosmopolitan in distribution. They can biomineralize intracellular ferrimagnetic crystals into organelles called magnetosomes and align along geomagnetic field lines[2]. Magnetotactic behaviour allows the enrichment and manipulation of magnetotactic bacteria from a microbial population using a magnet[3].

In this study, we employ high-throughput single-cell sequencing of naturally occurring consortia formed between magnetotactic bacteria and nanobacteria to unlock near-complete genomes from the symbiotic partners. Phylogenomic studies of the single-cell amplified genomes reveal the magnetotactic bacteria in association with the nanobacteria are from a single genus; the nanobacteria cover three different Patescibacteriota lineages that comprise two different taxonomic classes.

Single-cell amplified genomes from physically associated partners reveal metabolic repertoires in Patescibacteriota and their partners, enrichment in genes for type IV pili[4] and other cell to cell interaction genes, and perhaps signatures of metabolic complementarity with their magnetotactic hosts. The type IV pili machinery of the Patescibacteriota is different from the ones previously described in the literature.[5]

Our results demonstrate that single-cell consortia sequencing targeting magnetotactic bacteria is a powerful approach to access the physical relationships and the genomes of Patescibacteriota. These results highlight pili-mediated symbiosis as a key driver of their evolution and distribution in the environment. Access to high-quality Patescibacteriota genomes could also pave the way for co-cultures involving a broader range of organisms from the Microbial Dark Matter, thereby enabling a broader exploration of their biodiversity and ecology.

Unravelling hyperglycemia in diabetic cardiomyopathy: a multivalued computational approach

Poster

Baptiste Rivoirard [1], ***Sahar AGHAKHANI*** [1], ***Noura Nouali*** [2], ***Sébastien Medard*** [3], ***Asma Serier*** [1]

1. *AIXIAL Innovation Lab, AIXIAL SAS,* ***2.*** *Inserm UMR 1149, Centre de Recherche sur l'Inflammation,* ***3.*** *ALTEN*

Abstract

Diabetic cardiomyopathy (DbCM) primarily arises from a dysregulated glucose metabolism associated with diabetes mellitus leading to structural and functional heart anomalies. Despite growing interest in epigenetics lately, effective prevention and treatment strategies remain limited. Indeed, understanding the molecular mechanisms underlying this widespread complication of diabetes requires contextualizing implicated factors across multiple biological strata. Computational biology approaches, particularly through the reconstruction of literature- and data-driven static biological networks and their integrated dynamic simulation, may facilitate this detailed analysis. In this regard, our approach is based on the construction of the largest state-of-the-art theoretical static network of DbCM, compiling mechanistic evidence from diverse biological layers and sources. From this comprehensive knowledge base, we automatically inferred the most extensive multivalued models to date for DbCM, which was later parameterized using genomic data. This dynamic model underwent extensive validations and simulations, demonstrating its robustness and reliability. Our approach provides new insights into the complex interplay of factors driving DbCM by offering a robust framework to explore potential new therapeutic targets for DbCM or assess the potential impact of medications within this specific context as presented in some use-cases.

Unravelling the fine-scale genotype diversity and evolution of grapevine fanleaf virus

Poster

Sélim Ben Chéhida [1], ***Jeanne Juquel*** [2], ***Eva Chevalier*** [2], ***Pierre Mustin*** [2], ***Jean-Michel Hily*** [3], ***Wassim Rhalloussi*** [4], ***Carine Schmitt*** [4], ***Myriam Hagege*** [4], ***Isabelle Rachel Martin*** [3], ***Olivier Lemaire*** [4], ***Anne Sicard*** [4], ***Emmanuelle Vigne*** [4]

1. *INRAE, UMR 1131A Santé de la Vigne et Qualité du Vin,* ***2.*** *INRAE, Université de Strasbourg, UMR-A 1131 Santé de la Vigne et Qualité du Vin, 68000 Colmar, France,* ***3.*** *Institut Français de la Vigne et du Vin,* ***4.*** *INRAE, Université de Strasbourg, UMR-A 1131 Santé de la Vigne et Qualité du Vin*

Abstract

Grapevine fanleaf virus (GFLV; species *Nepovirus foliumflabelli*, family *Secoviridae*) is a major pathogen affecting vineyards worldwide, causing fanleaf degeneration disease generally resulting in significant economic losses. Its genome consists of two ssRNA(+) segments (RNA1 and RNA2). The virus is specifically transmitted in vineyards by the ectoparasitic nematode *Xiphinema index*, producing characteristic circular disease patches. Despite its long-standing characterization and global distribution, the fine-scale structure of GFLV populations remains poorly resolved, particularly at the patch scale.

Here, we present the investigation of GFLV diversity and evolution across six highly infected vineyard patches located in two major wine growing regions of France, Burgundy and Champagne. Leaves from one quarter of the grapevines in each patch were sampled, and total RNA was extracted followed by Illumina sequencing. After quality control, viral reads were *de novo* assembled for both genomic segments, yielding 389 RNA1 and 352 RNA2 consensus sequences. Maximum-likelihood phylogenies were inferred independently for RNA1 and RNA2 using IQ-TREE3.

To assess whether the observed diversity could be partitioned into robust genotypic clusters, GFLV genotypic diversity was explored using five delimitation models. Genotypic clusters were first delineated using a 5% nucleotide divergence cut-off based on both pairwise comparisons and maximum-likelihood phylogenetic distances. Phylogeny-based delimitation was performed using the multi-rate Poisson Tree Processes model (mPTP). Sequences were also partitioned using distance-based methods Automatic Barcode Gap Discovery (ABGD) and Assemble Species by Automatic Partitioning (ASAP). From the consensus of these five methods, 84 genotypes were identified for RNA1 and 93 for RNA2.

Whereas most genotypes were restricted to individual disease patches, indicating strong spatial structure, analyses of nucleotide diversity along the genome revealed consistent patterns across the patches. Together, these results provide a first view of the fine-scale genetic structure of GFLV populations within vineyards and will be used to investigate evolutionary processes shaping GFLV diversity.

Unveiling the dynamic transcriptome of the microsporidia parasite Anncaliia algerae during Human cell invasion.

Poster

Ivan Wawrzyniak [1], ***Reginald Akossi*** [2], ***Damien Courtine*** [1], ***Frédéric Delbac*** [1], ***Eric Peyretaillade*** [1]

1. *Laboratoire Microorganismes: Génome et Environnement, UMR 6023, CNRS, Université Clermont Auvergne, 63000 Clermont-Ferrand, France.,* ***2.*** *Trypanosome molecular biology unit, Plasmodium Chromatin and Transcription group, Parasites and Insect Vectors Department, Institut Pasteur Paris, 75015 Paris, France.*

Abstract

Anncaliia algerae is an emerging human pathogen and an obligate intracellular microsporidian parasite phylogenetically related to fungi. This species exhibits intriguing genomic features for a microsporidian, including a high proportion of transposable elements and repetitive sequences [1], as well as notable biological traits such as the ability to infect two distinct hosts, humans and anopheles mosquitoes [2]. A first investigation of host transcriptional response over a time course of infection of Human fibroblast Foreskin (HFF) cells at five time points (3h, 12h, 24h, 48h and 72h), using Illumina NovaSeq 6000 short-read sequencing, revealed extensive remodelling of host cellular processes during infection [3].

In this study, we investigate the parasite's own transcriptomic dynamics during infection of HFF cells. Principal component analysis (PCA) showed clear temporal separation of transcriptional profiles, consistent with progressive changes across the infection cycle. Differential expression analysis with DESeq2 identified sets of genes significantly up- or down-regulated at successive stages, while volcano plots illustrated the amplitude and significance of these transcriptional shifts. Gene Ontology (GO) enrichment analyses highlighted a temporal shift from pathway related to protein synthesis and proteostasis toward those involved in regulation, transport, signalling, and survival during parasite development.

Functional network analysis using STRING, combined with KEGG pathway enrichment, indicated modulation of major metabolic processes, including translation, nucleic acid metabolism and energy/redox metabolism, highlighting remodelling of the parasite's energetic capacity to sustain intracellular proliferation. Furthermore, clustering approach identified co-regulated gene clusters suggesting potential regulatory modules.

Together, these findings provide new insights into the transcriptional remodelling and adaptative strategies of *A. algerae* during human cell infection.

VacDesignR®: a computational tool to optimize viral-based individualized neoantigen therapeutic vaccine production

Poster

Anne-Isabelle Moro [1], ***Benoît Grellier*** [1]

1. *Transgene SA*

Abstract

Modified Vaccinia virus Ankara (MVA) is an attenuated vaccinia virus with a double-stranded DNA genome used to develop individualized neoantigen therapeutic vaccines (INTV), such as TG4050. TG4050 is currently evaluated in a Phase 1/2 clinical trial in Head and Neck cancers* (NCT04183166). During production, MVA are prone to homologous recombinant events; this genetic engineering advantage remains an issue if unwanted homologous regions are incorporated. Overexpression of hydrophobic peptides and peptides containing transmembrane (TM) domains, might impair the viability of cells and therefore the yield of virus produced. The successful incorporation of a high variety of peptides requires the control of these specific features by computational methods.

VacDesignR® is a software selecting and assembling up to 30 predicted immunogenic peptides (IPs) to be cloned in the MVA genome. The 3 main generation steps are the peptides selection, the inter-cassette repartition and the intra-cassette ordering. The IPs are filtered for homologous regions and sorted based on TM and Hydropathy (H) scores, defining 3 “production difficulties” classes: Low, Medium or High. IPs from the 3 classes are then dispatched to each cassette. This inter-cassette balanced repartition is performed 30,000 times to generate random batches and calculate H mean score. The intra-cassette repartition applies pre-defined slot positions and >2000 combinations are performed for each cassette. Only the fusions with the lowest H and TM scores are kept for final plasmid production.

Evaluated during the conception of MVA-based individualized neoantigen vectors, the production of recombinant viruses succeeded in <30% rate without optimization and >80% rate with VacDesignR®, which allows the use of this vector as INTV.

VacDesignR® is an efficient and effective computational tool to optimize plasmid design avoiding unwanted homologous recombination and production issues by keeping the TM and H scores low. Future versions including AI-based components are currently evaluated to improve performance.

URL

https://www.transgene.com/

Workflow development for automatically using Large Language Models to extract entities from a predefined corpus of scientific papers: creation of a knowledge graph for the fungal species Podospora anserina

Poster

Anakim Gualdoni [1], ***Pierre GROGNET*** [1], ***Fabienne MALAGNAC*** [1], ***Thomas Denecker*** [2], ***Gaelle LELANDAIS*** [1]

1. *I2BC, Université Paris-Saclay,* ***2.*** *IFB-core, Institut Français de Bioinformatique (IFB), CNRS, INSERM, INRAE, CEA, 94800 Villejuif, France.*

Abstract

Large Language Models (LLMs) are neural networks trained on massive amounts of text to predict the following word within a sequence, ensuring semantic coherence. Their significance in bioinformatics is rapidly growing. Notably, foundation models pre-trained on biological sequences are emerging for tasks such as structural prediction and functional annotation (1). Furthermore, LLM-driven automatic entity extraction now enables the analysis of unstructured data, including scientific literature. This capability relies on the attention mechanism in the Transformer architecture, which efficiently captures semantic regularities and contextual dependencies within texts (2).

The objective of this project is to leverage LLM-driven automatic entity extraction to construct knowledge graphs in bioinformatics. Here, we present preliminary results of implementing this approach for *Podospora anserina*, a fungal model organism in our laboratory. We started with a set of scientific articles curated by biological experts.

We developed an initial pipeline to automatically extract titles and authors from open-access papers indexed in Europe PMC, and map the extracted entities to a knowledge graph. This workflow uses DSPy, a programming framework that optimizes LLM outputs (3), CocoIndex (4), an Extract-Transform-Load framework (ETL) for LLMs enabling efficient creation and management of knowledge graphs in Neo4j, a graph database (5).

Subsequently, we generalized this pipeline to extract entities of interest from article abstracts without a predefined schema. This schema-free approach relies exclusively on the LLM's ability to assess term relevance based on a general context.

We evaluated the impact of key parameters, including prompt engineering strategies, corpus size, and computational resources, on the quality of the two generated knowledge graphs.

Authors Index

BGI Tech

Stereo-seq OMNI Service

Unparalleled insights at true single-cell resolution

Stereo-seq OMNI is a sequencing-based spatial multi-omics solution specifically optimized for archival Formalin-Fixed and Paraffin-Embedded (FFPE) tissues. It combines true single-cell level gene expression profiling with histology, empowering groundbreaking discoveries in clinical and translational research.

1. Gene expression profiling at true single-cell resolution, paired with histological analysis
2. Species-agnostic, and compatible with low-quality samples (DV200 ≥30%)
3. Random probe, designed for efficient capture of total RNA and microorganisms
4. Advanced bioinformatics service, with in-house software and designed pipeline

Please visit our booth to know more!

Discover More Services

Sequencing Services

- Single-Cell Sequencing
- Spatial Transcriptome
- Whole Genome Resequencing
- *De Novo* Sequencing
- Metagenome
- Transcriptome
- Whole Genome Bisulfite Sequencing

.....

Mass Spectrometry Services

- Quantitative Proteome
- Nanoproteome
- Metaproteome
- PTM Proteome
- Untargeted Metabolome
- High-Throughput Targeted Metabolome
- Lipidome

.....

Multi-Omics Services

- Single-Cell RNA-seq+ Spatial Transcriptome
- Single-Cell RNA-seq+ Bulk RNA-seq
- Transcriptome+Proteome+ Metabolome
- Metagenome/16S+Metabolome
- Dr. Tom Visualization System

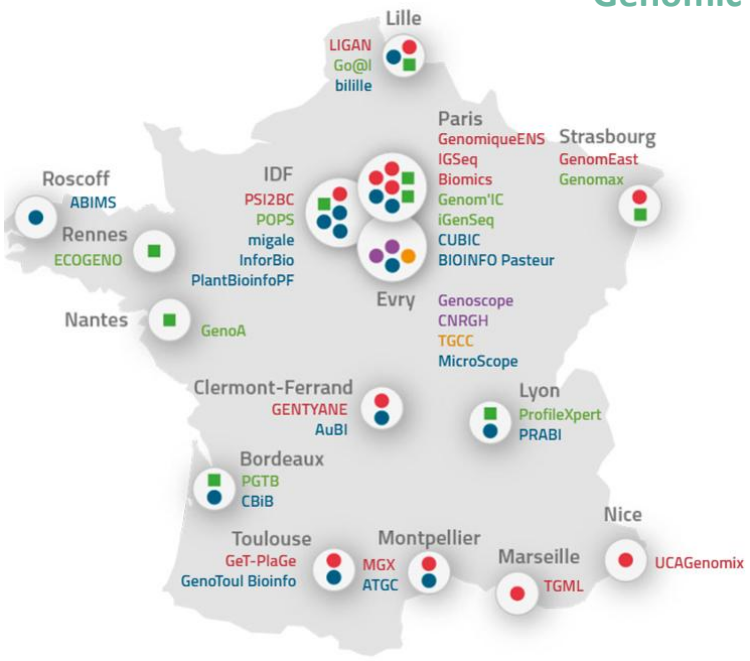
.....

Contact Us

info@bgi.com | www.bgi.com/global

FRANCE GENOMIQUE

A National World-class Infrastructure For Your Genomic Projects



- A national infrastructure created in 2012 thanks to a government grant from the "Investissements d'Avenir" program
- A coordinated network of the main French sequencing platforms

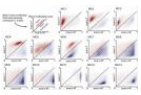
FRANCE GÉNOMIQUE MISSIONS

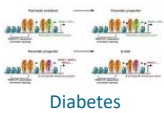
- Offer support for your projects
- Disseminate genomics knowledge
- Provide educational & professional training
- Provide state-of-the-art expertise in genomics & bioinformatics
- Give access to cutting edge technology & innovative approaches

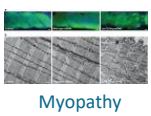
THE SEQUENCING RESOURCE

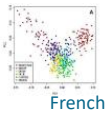
HIGH SCIENTIFIC IMPACT

- FG infrastructure helps the scientific community to produce high impact research through project funding, technology investment and development. Since 2012, more than 10k projects to which FG contributed led to > 2 500 publications.
- France Génomique directly funded 75 projects through 5 calls for proposals, which led to numerous high-quality publications in prestigious journals

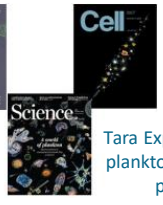

Liver Cancer


Diabetes


Myopathy



French Exome database



The Pea Genome

Tara Expeditions, plankton pole to pole


Ancient Horse

WHERE TO FIND US

www.france-genomique.org
contact@france-genomique.org
@fr-genomics.bsky.social
France Génomique


sfbi
société française de bioinformatique


cnrs


GDR
Groupement de recherche
BIMMM Bio-Informatique Moléculaire : Modélisation et Méthodologie


IFB
INSTITUT FRANÇAIS DE BIOINFORMATIQUE


BGI


FRANCE 2030


IA Grand Est
ENACT

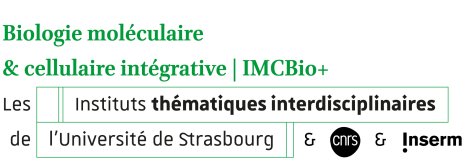

FRANCE GENOMIQUE


CANCÉROPÔLE Est
Régions Bourgogne-Franche-Comté / Grand Est


GREENOMICS
GENOMICS RESEARCH FOR ECO-EVOLUTION AND ENVIRONMENT
sfé²


iCUBE


IMC Bio


Biologie moléculaire
& cellulaire intégrative | IMCBio+
Les Instituts thématiques interdisciplinaires
de l'Université de Strasbourg & cnrs & Inserm


INRAE
ALIMH


INRAE
BAP


INRAE
GA


INRAE
PHASE


INRAE
SPE


INRAE
SA


Inria


Loria


INCA


RNA networks
Net RNA


Strasbourg.eu
eurometropole


transgene


Université
de Strasbourg


IBMP
Institut de biologie moléculaire des plantes


igbmc
institut de génétique et de biologie moléculaire et cellulaire


INCI


gm
Génétique Moléculaire = Génomique = Microbiologie


SVQV
Santé de la Vigne & Qualité du Vin