



Proceedings

Preface

Dear Participants and Colleagues,

Welcome to the 24rd JOBIM edition, that is for the first time implemented in a multi-site configuration. This achievement would not have been possible without the tremendous efforts and dedication of the organizing committee, who did an impressive work to ensure a great conference experience. We would like to warmly thank them and the local committees of Nancy, Nice, Pointe-à-Pitre, Plouzané and Tours sites for their enthusiasm and commitment to making this conference a success.

This edition regroups more than 350 participants among which many submitted a communication. We extend our thanks to the whole program committee for their invaluable contribution in the selection process, which led to the selection of 24 oral communications, 122 posters (including 17 selected for flash talks). We also acknowledge all the authors for their valuable contributions and congratulate those whose work was selected for presentation.

Additionally, the program committee also selected 2 mini-symposium and 4 keynotes. Mini-symposiums are dedicated to multi-omics integration in Nice and Metagenomics in Pointe-à-Pitre. They will be broadcast across all JOBIM sites and remotely.

We are particularly honored to host four esteemed keynote speakers: Allison Ballandras-Colas, Anaïs Baudot, Carl Herrmann, and Rayan Chikhi. Their insights and expertise have added immense value to the conference and have provided a unique perspective on the latest advancements in our field. We are grateful to each of them for accepting our invitation and for their enlightening presentations.

We would like to acknowledge the support of our institutional partners, namely the IFB, GdR GE, GdR BIM and SFBI. Their support and partnership have been instrumental in making this conference possible, and we are truly grateful for their continuous commitment to fostering scientific collaboration and advancement.

Finally, we would like to thank everyone involved in making the JOBIM conference a warm place for fruitful exchanges around computational biology.

Enjoy the proceedings, and we look forward to your continued participation and engagement in future editions of JOBIM.

Matthias Zytnecki & Delphine Potier

Presidents of the Organization Committee

Corre	Erwan	CNRS	Roscoff
Morin	Emmanuelle	INRAE	Nancy
Noël	Cyril	IFREMER	Plouzané

Members of the Organization Committee

		Nancy	
Morin	Emmanuelle	INRAE	Nancy
Malliavin	Thérèse	Université de Lorraine	Nancy
Auer	Lucas	INRAE	Nancy
Corre	Emma	INRAE	Nancy
Kreplak	Jonathan	INRAE	Dijon
		Nice	
Robbe-Sermesant	Karine	INRAE	Nice
Bécavin	Christophe	CNRS	Nice
Bottini	Silvia	Université Côte d'Azur	Nice
Croce	Olivier	CNRS	Nice
Gautier	Romain	CNRS	Sophia Antipolis
Sarti	Edoardo	INRIA	Sophia Antipolis
Da Rocha	Martine	INRAE	Sophia Antipolis
Dagnino	Sonia	Université Côte d'Azur	Nice
Bailly-Bechet	Marc	INRAE	Nice
		Pointe-à-Pitre	
Couvin	David	Institut Pasteur Guadeloupe	Pointe-à-Pitre
Quétel	Isaure	Institut Pasteur Guadeloupe	Pointe-à-Pitre
Stattner	Erick	Université des Antilles	Pointe-à-Pitre
Segretier	Wilfried	Université des Antilles	Pointe-à-Pitre
Pasquier	Raphael	Université des Antilles	Pointe-à-Pitre
Gonzalez-Rizzo	Silvina	Université des Antilles	Pointe-à-Pitre
Bambou	Jean-Christophe	INRAE	Petit-Bourg
Bhakkan	Bernard	CHU Guadeloupe	Pointe-à-Pitre
Gaete	Stanie	Karubiotec, CHU Guadeloupe	Pointe-à-Pitre
Doncescu	Andrei	Université des Antilles	Pointe-à-Pitre
		Plouzané	
Corre	Erwan	CNRS	Roscoff
Le Corguillé	Gildas	Sorbonne Université	Roscoff
Rousseau	Coralie	CNRS	Roscoff
Durand	Patrick	IFREMER	Plouzané
Hellec	Elisabeth	IFREMER	Plouzané
Leroi	Laura	IFREMER	Plouzané
Noël	Cyril	IFREMER	Plouzané
		Tours	
Lefort	Gaëlle	INRAE	Tours
Rogier	Odile	INRAE	Orléans

Presidents of the Programme Committee

Potier	Delphine	CNRS	SFBI
Zytnicki	Matthias	INRAE	SFBI

Members of the Programme Committee

Abby	Sophie	CNRS	
Aubert	Julie	INRAE	
Ballester	Benoît	INSERM	
Baudot	Anaïs	CNRS	
Bérard	Sèverine	Université de Montpellier	
Blum	Yuna	CNRS	
Bourdon	Jérémie	CNRS	
Brun	Christine	CNRS	
Camproux	Anne-Claude	Université Paris-Diderot	
Cazals	Frédéric	Inria	
Clément	Yves	Institut Jacques Monod	SFBI
Corre	Erwan	CNRS	SFBI
Dameron	Olivier	Université de Rennes	
Djebali	Sarah	INSERM	
Durand	Patrick	IFREMER	
Eveillard	Damien	CNRS	
Fiston-Lavier	Anna-Sophie	Université de Montpellier	SFBI
Labesse	Gilles	CNRS	
Lacroix	Vincent	Université de Lyon	
Lagarrigue	Sandrine	INRAE	
Lecellier	Charles	CNRS	
Lerat	Emmanuelle	CNRS	
Loux	Valentin	INRAE	IFB
Mariadassou	Mahendra	INRAE	GdR BIM
Neuvial	Pierre	CNRS	
Nickolski	Macha	CNRS	IFB
Pécreaux	Jacques	CNRS	
Pelletier	Eric	CEA	
Peterlongo	Pierre	Inria	
Ponty	Yann	CNRS	
Rivals	Eric	CNRS	GdR BIM
Roest-Crollius	Hugues	CNRS	
Salson	Mikaël	Université de Lille	
Scornavacca	Céline	CNRS	
Siegel	Anne	CNRS	
Thébault	Patricia	Université de Bordeaux	
Théret	Nathalie	INSERM	
Uricaru	Raluca	Université de Bordeaux	
Vialaneix	Nathalie	INRAE	

List of Presentations

Keynote speakers	13
Allison Ballandras-Colas: Catalytic integration, the point of no return during retroviral infection	14
Anais Baudot: Network-based data integration for genetic diseases	15
Rayan Chikhi: Living in the Future of Sequence Bioinformatics	16
Carl Herrmann: Explainable machine-learning models for omics: from black to grey boxes . . .	17
Mini symposium 1: Metagenomic analyses in an island context (Guadeloupe) and in French Guiana	19
Bacterial microbiota management in free-living amoebae (Heterolobosea) isolated from water: the impact of amoebae identity, grazing conditions and passage number [Quétel, Isaure] . .	20
Resistome and microbiome of biofilms [Batantou, Degrâce]	21
Influence of breeding sites on Aedes aegypti microbiota and vectorial capacity [Vega-Rua, Anubis]	22
Larval diet influence Aedes aegypti microbiota and fitness in laboratory and natural rearing conditions [Calvez, Elodie]	23
Impact of amoxicillin/clavulanic acid on Aedes aegypti microbiota and its capacity to transmit dengue virus [Garcia-Van Smévoorde, Margot]	24
Disentangle chimeric (d-chimer) sequences in de novo assembled viromes [Tirera, Sourakhata] .	25
Mini symposium 2: Multi-omics integration: challenges and perspectives	27
Graph-based multi-omics integrative approaches to study microbial ecosystems across scales [Chaffron, Samuel]	28
Multi-omic data integration with prior knowledge to decipher signaling and metabolic deregulation in complex diseases [Dugourd, Aurélien]	29
Imaging-genetic approaches to study the human brain [Philippe, Cathy]	30
Multi-omics data integration methods: kernel and other machine learning approaches [Vialaneix, Nathalie]	31
Session 1: Structural bioinformatics and proteomics / Systems biology and metabolomic	33
The Martini Database (MAD): a web service to facilitate molecular simulations (platform) [Marin, Romuald]	34
On the predictability of A-minor motifs from their local contexts (highlight) [Gianfrotta, Coline]	36
An agnostic analysis of the human AlphaFold2 proteome using local protein conformations (highlight) [De Brevern, Alexandre]	37
An Agent-Based Model of Monocyte Differentiation into Tumor-Associated Macrophages in Chronic Lymphocytic Leukemia (highlight) [Verstraete, Nina]	38
Multi-omics characterization of Richter syndrome unlocks classifiers and predictors of outcome into broader and heterogeneous lymphoma datasets (highlight) [Hergalant, Sébastien] . . .	39
Session 2: Evolution	41
Horizontal transfer shapes the subsequent emergence of antibiotic resistance by point mutations (proceedings) [Coluzzi, Charles]	42
Limited Transmission of Klebsiella pneumoniae among Humans, Animals, and the Environment in a Caribbean Island, Guadeloupe (French West Indies) (highlight) [Breurec, Sebastien] .	50
Metagenome reveals caprine abomasal microbiota diversity at early and late stages of Haemonchus contortus infection. (highlight) [Bambou, Jean-Christophe]	65
KaruBioNet: a network and discussion group for a better collaboration and structuring of bioinformatics in Guadeloupe (French West Indies) (platform) [Couvin, David]	66

SeBiMER: the bioinformatics platform of Ifremer (platform) [Durand, Patrick]	68
Session 3: Statistics, machine learning, artificial intelligence and image analysis	71
DeCovarT: Robust deconvolution of cell mixture in transcriptomic samples by leveraging cross-talk between genes (proceedings) [Chassagnol, Bastien]	72
Assessing goats' fecal avoidance using image analysis based monitoring (highlight) [Bonneau, Mathieu]	81
Multivariate Analysis of RNA Chemistry Marks Uncovers Epitranscriptomics-Based Biomarker Signature for Adult Diffuse Glioma Diagnostics (highlight) [Rivals, Eric]	82
LEAF: a machine learning approach to predict effector proteins in Candidatus Phytoplasma (proceedings) [Calia, Giulia]	83
Goodness of Fit for Bayesian Generative Models with Applications in Population Genetics (proceedings) [Le Mailloux, Guillaume]	90
Session 4: Functional and integrative genomics	99
Alteration of ribosome function upon 5-fluorouracil treatment favors cancer cell drug-tolerance (highlight) [Ripoll, Julie]	100
Comparative analysis of whole blood transcriptomics between European and local Caribbean pigs in response to feed restriction in a tropical climate (proceedings) [Poullet, Nausicaa]	101
Drug effects in gene regulation: how multi-omics integration can benefit gene therapy design (platform) [Cherchame, Emeline]	109
AnnotSV and knotAnnotSV: a webserver for human structural variations annotations and analysis (platform) [Geoffroy, Véronique]	111
Cirscan: a shiny application to decipher circRNA-miRNA-mRNA networks and condition-specific sponge mechanisms. (proceedings) [Fraboulet, Rose-Marie]	113
Session 5: Algorithms and data structures for sequences / Knowledge representation	123
HairSplitter: separating strains in metagenome assemblies with long reads (proceedings) [Faure, Roland]	124
Opening the Black Box of Imputation Software to Study the Impact of Reference Panel Composition on Performance (highlight) [Herzig, Anthony]	132
Session 6: Workflows, reproducibility and open science	133
SnakeMAGs: a simple, efficient, flexible and scalable workflow to reconstruct prokaryotic genomes from metagenomes (highlight) [Hervé, Vincent]	134
BEAURIS: an automated, modular, and FAIR system for large-scale genome data management (platform) [Boudet, Matéo]	136
FlashTalk 1	139
Compositional biases promoting self-assembly establish a link between the genome- and the cell-spatial self-organization [Lapendry, Audrey]	140
DLScalf : Deep Learning and Hi-C data for chimeric contigs detection [Mergez, Alexis]	141
Exploring the Potential of a Biomimetic Fibronectin Motif to Interact with Type I Collagen for Tissue Regeneration: In Silico and Experimental Analyses [Eid, Jad]	142
Genes encoding teleost orthologs of human signal transduction proteins remain in duplicate or in triplicate more frequently than the whole genome [Picolo, Floriane]	144
Handling confounding factors in analyzing the transcriptomic data from Chornobyl tree frogs [Goujon, Elen]	146
THEMA: Identification of molecular mechanisms by which Tumor HEterogeneity influences disease outcome: high-dimension Mediation Analysis to link causes and consequences [Pit-tion, Florence]	147
Recent introduction of <i>Angiostrongylus cantonensis</i> and its intermediate host <i>Achatina fulica</i> in Guadeloupe evidenced by phylogenetic analyses [Gamiette, Gélixa]	148
End of the beginning: telomeres are only at one side of the chromosomes in the nematode <i>Meloidogyne incognita</i> [Robbe-Sermesant, Karine]	149
CellFromSpace: A versatile tool for spatial transcriptomic data analysis through reference-free deconvolution and guided cell type/activity annotation [Thuilliez, Corentin]	150
Spatial methods for the analysis of genetic data [Guivarch, Maël]	151

FlashTalk 2	153
GRUPS-rs - a high-performance ancient DNA genetic relatedness estimation software relying on pedigree simulations [Lefeuvre, Maël]	154
L'impact des outils d'assemblage sur le typage des pathogènes bactériens [Merda, Déborah] . .	155
Modelling of 3D structures of aminoacylases from <i>Streptomyces ambofaciens</i> and molecular rules for their specificity and their regioselectivity in lysine N-acylation [Genesseaux, Laureline]	157
Inferring and comparing metabolism accross heterogeneous sets of annotated genomes using AuCoMe [Markov, Gabriel]	158
Multiplex Network Exploration to define The landscape of Premature Aging Diseases [Beust, Cécile]	159
PainterPipe: a pipeline for genetic variant fine-mapping using functional annotations [Gerber, Zoé]	160
Exploitation des métadonnées et données génomiques d'agents pathogènes dans l'outil Bac'PACK pour améliorer les approches One HEALTH [Chesnais, Virginie]	161
The flies' route of 3GCR <i>E. coli</i> dissemination in beef cattle farm: from ecosystem to molecular scale [Ferdinand, Séverine]	162
Poster 1	171
Alteration of transposable element mutation rate by nucleosome positioning [Sassolas, Fabien] .	172
Challenges of genomic data generation for non-model complex species [Doré, Guillaume] . . .	173
Characterization of Transposable Elements in Pangenomes [Saidi, Somia]	174
Evolution and molecular characterization of internal targeting signals in mitochondrial proteins [Galan, Clément]	175
Evolution of corn metabolites detoxification in <i>Ostrinia</i> [El Khatib, Mariam]	176
Evolution of habenular asymmetries in gnathostomes: a transcriptomic approach [Mayeur, Hélène]	177
Genomics landscape shaped by transposable elements through rust pathogens history [Corre, Emma]	178
Highlighting the evolution towards antibiotic cross-resistance in <i>E. coli</i> biofilms exposed to biocides using large-scale comparative genomics [Lemée, Pierre]	179
La phylogénie pour le suivi épidémiologique de la fièvre catarrhale en Guyane [Merda, Déborah]	180
Testing pangenomic tools for structural variant detection in non-model organisms [Denni, Sukanya]	182
The complete mitochondrial genome of coffee leafminer <i>Leucoptera coffeella</i> , a major pest of coffee crops [Pereira Dos Santos, Mateus]	183
Free-living forms and EVEs of brown algae viruses [Massau, Karine]	184
Insect invaders analysis pipeline - What happens to the genetic load when alien species become invasive? [Porro, Barbara]	186
phylEntropy, a web-based tool for various data visualization applications [Cazenave, Damien] .	188
Poster 2	189
Benchmarking read mapping on pangenomic variation graphs [Bouamout, Hajar]	190
Development of a supervised ctDNA analysis pipeline to improve minimal residual disease monitoring in lymphoma [Gomes, Lucie]	191
P-GRE: a new fully-automatic pipeline for the precise annotation of the structure and position of pseudogenes [Cabanac, Sébastien]	192
Visual Representation of Genomes [Stattner, Erick]	193
BeeDeeM: a general-purpose bioinformatics databank manager system [Durand, Patrick]	194
Characterization of particulate matter in the Caribbean area [Euphrasie-Clotilde, Lovely]	195
Equine individual limits for monitoring IGF-1 levels in plasma: implementation to the Equine Biological Passport [Barnabé, Agnès]	196
Exploring Immune and Tumor Cells in Gliomas Highly Infiltrated by Lymphocytes through Single-Cell RNA-seq and Single-Cell CITE-seq Data Analysis [Brocic, Jovana]	197
FliesDB a long story of genomic interface [Samson, Franck]	198
GenoFig: A user-friendly application for the visualization and comparison of genomic regions [Leclercq, Sébastien]	199
PanExplorer: a web application for exploratory analysis and visualization of microbial pangenomes [Meyer, Damien]	200
ToolDirectory: Dynamic visualization of softwares managed by Bioinformatics Core Facilities [Cormier, Alexandre]	201
A multi-block approach to improve deconvolution of cancer omic data [Amblard, Elise]	202

ABEILLE: a novel method for ABerrant Ex-pression Identification empLoying machine Learning from RNA-sequencing data [Labory, Justine]	203
adverSCarial: a tool for evaluating adversarial attacks on single-cell transcriptomics classifiers [Fiévet, Ghislain]	204
An automated PVC/SR QRS discriminator on 12-lead ECG [Ahraoui, Amèle]	205
ECGtizer: an open-source, fully automated and versatile ECG digitisation tool that generates ready to use XML data for AI-based annotations and analyses. [Lence, Alex]	206
Inference of biological interactions on large heterogenous graph networks with machine learning [Toffano, Antoine]	207
Issues in UK Biobank for GWAS: case-control definition and imbalance [Derouin, Margot]	208
Prediction and classification of methylomic class of brain CNS tumour [Jossaud, Fabien]	209
Random Walk with Restart on multilayer networks: from node prioritization to supervised link prediction and beyond [Briere, Galadriel]	210
Scoring and ranking strategies to benchmark cell type deconvolution pipelines [Bertrand, Vadim]	211
Development of a knowledge graph framework to ease and empower translational approaches in plant research: a use-case study on grain legumes [Imbert, Baptiste]	212
Poster 3	213
A 4-Year retrospective study of the presence of thermophilic free-living amoebae in recreative baths in Guadeloupe [Quétel, Isaure]	214
A novel and dual digestive symbiosis scales up the nutrition and immune system of the holobiont <i>Rimicaris exoculata</i> [Aubé, Johanne]	216
Assessing the hidden biodiversity of coral reefs in Guadeloupe using Autonomous Reef Monitoring Structures (ARMS) [Bezault, Etienne]	217
Characterization and quantification of antibiotic resistance gene variants in gut microbiota [Sidibé, Ouléye]	218
Characterization of the functional composition of the Human Gut Microbiome in Liver Cirrhosis and Colorectal Cancer (CRC) and identification of candidate biomarkers using AI [Henecart, Baptiste]	219
Deciphering and cataloging the genomic and functional diversity of the French cheese microbiota [Gardon, Hélène]	220
Identification of a microbiome, the advantages of the metagenomic method over the classical 16S method [Peticca, Aurélie]	221
Microbial diversity of <i>Beggiatoa</i> mats reveal a different taxonomic profile from marine mangrove sediments in urban and natural sites of the Caribbean [Martinez-Noriega, Mariana]	222
Suitability of Nanopore adaptative sampling for metabarcoding approaches. Is it possible to directly remove chloroplast sequences from algal samples during sequencing? [Rousseau, Coralie]	224
Tracing human-borne bacterial contaminants from wastewater treatment plants to coral reefs in the Caribbean Sea [Urrutia, Ander]	225
A global catalogue of genomes and protein sequences from the termite microbiome [Tadrent, Nachida]	226
ABRomics: An integrated multi-omics platform for antibiotic resistance research and public health [Médigue, Claudine]	227
Automatic annotation using RNA-seq data [Ziane, Khaoula]	229
Bioinformatics challenges in the analysis of gastruloid time-series single-cell RNA-seq data [Chevalier, Céline]	230
Double approche, expérimentale et bio-informatique pour l'étude de l'épissage des ARN pré-messagers dans les cancers MSI [Kon-Sun-Tack, Fabien]	231
Evaluation of Oxford Nanopore R9.4.1/Kit10 , R10.4/Kit12 and R10.4.1/Kit14 sequencing for minority variants analysis [Boyer, Théophile]	232
External quality assessment of SARS-CoV-2 variants identification and whole genome sequencing across 45 French laboratories [Tonazzolli, Arianna]	233
Functional annotation of dinoflagellate protein sequences and structural prediction by deep learning approach [Rousseau, Jérémy]	235
Genome analysis of SNP and SV in the admixed Creole cattle of Guadeloupe reveals new adaptive mechanisms to tropical production system [Naves, Michel]	236
Genome-scale essential gene discovery in a bacterium by hyper saturated transposon insertion amplicon sequencing [Jarrige, Domitille]	237

Genome-wide CRISPR screens in B cell lymphomas reveal novel oncogenic dependencies [Soun, Camille]	238
Identifying genetic factors influencing the development of vascular aneurysms in Autosomal Dominant Polycystic Kidney Disease [Lemoine, Hugo]	239
MicroRNA annotation tool comparison in animal genomes [Racoupeau, Martin]	240
Poster 4	241
A novel hierarchical agglomerative clustering algorithm for inferring a representative set of 3D RNA conformations [Chauvot De Beauchêne, Isaure]	242
B-cell epitope prediction on HLA antigens using molecular dynamics simulation data [Amaya, Diego]	243
Comparative analysis of topologically associating domains -TAD- callers [Rique, Flavian]	244
Evaluation of SARS-CoV-2 Spike RBD interactions with Angiotensin Converting Enzyme 2 of multiple species using Molecular Dynamics Simulations [Garcia, Damien]	245
Influence of IDR deamidations on Bcl-xL structure and function [Hunault, César]	246
Using bulk RNA-seq and iClip-seq analysis to investigate mRNA localization in synapses and uncover the role of the IMP protein [Laghrissi, Hiba]	247
Enzymatic network programming in non-living biomachines for medical diagnosis [Davy, Martin]	248
Inferring and comparing metabolism accross heterogeneous sets of annotated genomes using Au-CoMe [Markov, Gabriel]	249
Metabolomic Modeling of Microbial Interactions for Enhanced Hydrogen Production [Marbehan, Xavier]	250
Mise en place d'un outil analytique en Guadeloupe pour le dosage de la chlordécone dans le sérum humain [Couvin, David]	251
ODAMNet: a Python package to identify molecular relationships between chemicals and rare diseases using overlap, active module or random walk approaches [Térézol, Morgane]	252
Study of the anaerobic degradability of chlordecone by mixed cultures [Gruel, Gaëlle]	253
Poster 5	255
Exploring the epigenetic regulation of alternative splicing in the context of mouse spermatogenesis [Feudjio, Olivier]	256
FAIR principles and open science in marine passive acoustic monitoring with the OSmOSE toolkit [Loire, Benjamin]	257
Impact of formalin on WGS PCR Free data reliability and accuracy from FFPE tumor tissue samples in a clinical context for the French Genomic Medicine Plan 2025 [Rouillon, Marine]	258
ISO-seq transcriptomics analysis for polyploid genomes [Pochon, Mathis]	259
QuaDS: Qualitative-Quantitative Descriptive Statistics [Bouanich, Andréa]	260
sRNA-pipe: a Nextflow-based pipeline for small RNA analysis [Pham, Hoang-Giang]	261
KiNext: A pipeline for the identification and classification of protein kinases [Hellec, Elisabeth]	262
AskoR, an R package for easy RNA-Seq data analysis [Gazengel, Kévin]	263
Method optimization for Rapid Pathogen Identification in Lower Respiratory Infection in Low Resource Settings [Arnaud, Jean]	264
Integrating the miRnome into multiple omics of Richter Transformation: insights into the development of aggressive lymphomas [Piucco, Romain]	266
Poster 6	267
The Carbohydrate-Active enZyme database: literature, functions, subfamilies and recent Python scripts to update it [Boulinguez, Matthieu]	268
BIPAA, Bioinformatics Platform for the Agroecosystems Arthropods [Robin, Stéphanie]	270
MoPSeq-DB: the reference database and visualisation platform for marine mollusc pathogens genomes [Battistel, Clémentine]	272
InDeepNet a web application to assist drug design [Mareuil, Fabien]	274
ABiMS: Analysis and Bioinformatics for Marine Science [Corre, Erwan]	276
Joint transcriptome and translatoome analysis: a reproducible pipeline [Ripoll, Julie]	277
MOAL: Improving the Reproducibility of OMIC Bioanalysis [Dumont, Florent]	278
The Reference, Innovation, Expertise and Transfer Center (CReFIX) of the French Genomic Medicine Plan 2025 [Letexier, Mélanie]	280
ETBII: a new IFB school on Integrative Bioinformatics [Khamvongsa-Charbonnier, Lucie]	281

DeepOmics Submission, a plug-in tool to facilitate the submission of meta-omics data to the ENA [Rousseau, Baptiste]	282
athENA: a Nextflow pipeline for sequencing data brokering to ENA [Auffret, Pauline]	283
Epitranscriptomic analyses by Nanopore direct RNA sequencing at the I2BC next-generation sequencing facility [Ouazahrou, Rania]	284
PredomicsApp: R shiny web application for interpretable and accurate construction of prediction models for OMICs data [Roy, Gaspar]	285
The Rhodexplorer Genome Database: a Multi-Scale Genomic and Transcriptomic Data Resource for the Red Algae [Brillet-Guéguen, Loraine]	286
The Migale bioinformatics core facility [Loux, Valentin]	287
L'institut Français de Bioinformatique: Centre de Référence Thématique Biologie-Santé dans l'écosystème Recherche Data Gouv [Devignes, Marie-Dominique]	288
RGB: the Guadeloupean Network of CRBs [Gaete, Stanie]	289
MERIT : réseau MetiER en bioinformaTique [Bihouée, Audrey]	290
Ferments du Futur : une plateforme unique en Europe qui entend accélérer la recherche et l'innovation sur les ferments, les aliments fermentés et la biopréservation dans les 10 prochaines années [Schbath, Sophie]	291
Building Research Capacity for Diagnostics, Exposome, and Bioinformatics in the Caribbean: A Collaborative Thesis Project [Arnaud, Jean]	292

Keynote speakers

Catalytic integration, the point of no return during retroviral infection

Allison ALLISON BALLANDRAS-COLAS¹

¹ Institute of Structural Biology (IBS), Grenoble, France

Corresponding Author: Allison.Ballandras-Colas@ibs.fr

Retroviruses are RNA viruses that retrotranscribe their genome into double-stranded DNA and then insert it into the DNA of the infected cell. This stage, called integration, marks the point of no return in the retroviral cycle, and is catalyzed by the retroviral integrase protein. Thus, the human immunodeficiency virus (HIV) integrase protein, belonging to the *lentivirus* genus of the *Retroviridae* family, is an important therapeutic target.

In this retrospective, I present cryo-EM structures of integrase proteins from different genus of retroviruses, and how such structures provide a better understanding of the molecular mechanism of retroviral integration, and of the inhibition by molecules used in anti-HIV therapy.

Network-based data integration for genetic diseases

Anaïs BAUDOT¹

¹ Marseille Medical Genetics (MMG), Marseille, France

Corresponding Author: anaïs.baudot@univ-amu.fr

Recent technological advances and the growing availability of biomedical datasets offer unprecedented opportunities to better understand human diseases. However, translating the sheer volume and heterogeneity of these data into meaningful insights require proper computational strategies. In this talk, I will present different network-based approaches for the integration of heterogeneous datasets. I will describe multilayer networks that incorporate different sources of biomedical interactions, as well as associated network exploration algorithms. I will illustrate the application of these different algorithms in the context of the analysis of rare genetic diseases, which raise various challenges: many patients are undiagnosed, phenotypes can be highly heterogeneous, and only a few treatments exist.

Living in the Future of Sequence Bioinformatics

Rayan CHIKHI¹

¹ Institut Pasteur, Paris, France

Corresponding Author: rayan.chikhi@pasteur.fr

This talk will put a spotlight on a gap that currently exists between the conventional way genomic sequencing data is analyzed, using single machines, clusters, standard workflows, etc.. and recent cutting-edge big data analyses using cloud resources. Mainstream genomic projects generally jointly analyze a handful, up to hundreds, or maybe even thousands, of whole-genome sequencing datasets. Many efforts have been carried by biologists and computer scientists to optimize methods, software tools, and pipelines, to reach that goal and maximize the potential for new discoveries. However, we will soon be faced with analyses that will require to tackle millions of complete genomes, and neither our current methods nor computational systems are not ready for that scale.

In 2020, in the midst of the Covid crisis, I took part in a consortium that performed one of the largest bioinformatics analyses to date. We performed petabase-scale alignment and assembly of all RNA-sequencing data and discovered new coronavirus species, as well as an order of magnitude more RNA virus species than previously known. This particular project will not be the focus of the talk, but, using the experience gained from this large-scale analysis I will give a few tricks and perspectives on what future genomics analyses could look like. We will talk about accessing big genomic resources efficiently, and consider the use-case of rapid sequence alignment.

Explainable machine-learning models for omics: from black to grey boxes

Carl HERRMANN¹

¹ BioQuant Center & Medical Faculty Heidelberg, Heidelberg, Germany

Corresponding Author: carl.herrmann@bioquant.uni-heidelberg.de

Deep-learning approaches are increasingly being adopted in biomedical application such as image analysis or genomics, in particular single-cell genomics. These methods have numerous advantages, such as capturing signal at different scale and modeling complex non-linear behavior. However, the prize to pay is a loss in interpretability of the models. Hence, an increasing effort is driven towards building explainable AI models in biomedecine, in which the impact of input features to model performance can be extracted or which are intrinsically intrinsically interpretable. We will review the challenges of building interpretable AI models, and present several recent approaches to map prior biological information to deep-learning models with applications in genomics.

Mini symposium 1: Metagenomic analyses in an island context (Guadeloupe) and in French Guiana

Organization

David Couvin, Isaure Quénel

Bacterial microbiota management in free-living amoebae (Heterolobosea) isolated from water: the impact of amoebae identity, grazing conditions and passage number

Isaure QUETEL¹

¹ Institut Pasteur de Guadeloupe, Pointe-à-Pitre, Guadeloupe, France

Corresponding Author: iquetel@pasteur-guadeloupe.fr

Free-living amoebae (FLA) are ubiquitous protists found in soil and freshwater, mainly feeding on bacteria. They are also well-known reservoirs and vectors for the transmission of amoeba-resistant bacteria (ARB), most of which are pathogenic to humans. Yet, the natural microbiome of wild amoebae remains largely unknown and the effects of amoeba identity and environmental factors on bacterial microbiome composition unexplored.

Monocultures of *Naegleria australiensis*, *Naegleria KDN1*, *Paravahlkampfia ustiana* and *Vermamoeba vermiformis* isolated from recreational waters in Guadeloupe were passaged with or without different food sources (*E. coli*, yeast, fetal calf serum and water) during successive replication cycles. The whole bacterial microbiome of the recreational baths and the amoebae was characterized using 16S rRNA metabarcoding. The culturable subset of amoebae-associated bacteria was analyzed by mass spectrometry (MALDI-TOF MS) and disk diffusion method to assess bacterial antibiotic resistance. Transmission electron microscopy allowed to locate the bacteria inside the amoebae cysts.

Resistome and microbiome of biofilms

Degrâce BATANTOU¹

¹ Institut Pasteur de Guadeloupe, Pointe-à-Pitre, Guadeloupe, France

Corresponding Author: dbatantou@pasteur-guadeloupe.fr

Guadeloupe, a French overseas territory located in the Caribbean, is a very high resource country according to the Human Development Index in 2013 (<http://hdr.undp.org/>). Also called "The island of beautiful waters" because of the many rivers and waterfalls that border the southern part of the island, Guadeloupe is confronted with the pollution of surface water and groundwater. This water pollution can result from the influx of chemical and biological factors from local treatment plants as well as industrial and agricultural effluents (http://www.comite-de-bassin-guadeloupe.fr/gestion_sources-contamination.php). Contaminated waterways are known spreaders of antibiotic resistance. Studies have shown that humans can be exposed to potentially pathogenic and antibiotic-resistant strains of enterobacteriaceae including *Escherichia coli* (*E. coli*) during recreational activities around the world (Leonard et al., 2018; Fernandes et al., 2017). On the island, bathing sites are identified and are subject to periodic health checks carried out by the Pasteur Institute of Guadeloupe (IPG) mandated by the Regional Health Agency (ARS) of Guadeloupe. In 2020, 145 bathing sites were subject to health monitoring, i.e. 1,465 samples. These data showed that 267 samples had an *Escherichia coli* rate greater than 100 MPN per 100 mL, indicating potential exposure to faecal coliforms that may be resistant to antibiotics. A previous study by Guyomard et al confirmed the presence of these resistant germs in the rivers of Guadeloupe.

Thus the main objective of this study is to estimate the proportion of bathing areas with ESBL *Escherichia coli* and to quantify their rate in the recreational waters of Guadeloupe. Secondary objectives are to use whole genome sequencing to identify the genetic background of susceptible and resistant isolates of *E. coli* collected from bathing waters, to conceptualize a method for collecting biofilms at river level and to study the microbiome and the resistome present in bathing waters on a subset of water samples from recreational aquatic areas.

Influence of breeding sites on *Aedes aegypti* microbiota and vectorial capacity

Elodie CALVEZ¹

¹ Institut Pasteur de Guadeloupe, Pointe-à-Pitre, Guadeloupe, France

Corresponding Author: ecalvez@pasteur-guadeloupe.fr

Mosquito microbiota and development could be influenced by the presence of bacteria and nutrients in the rearing water. Here, we studied the influence of four diets, commonly used in laboratory, on *Aedes aegypti* microbiota and fitness in both laboratory and field water conditions. Furthermore, we evaluated if these diets could modulate the transmission of dengue virus by this vector.

Larval diet influence *Aedes aegypti* microbiota and fitness in laboratory and natural rearing conditions

Anubis VEGA-RUA¹

¹ Laboratory of Vector Control Research, Unit Transmission Reservoir and Pathogens Diversity, Institut Pasteur de la Guadeloupe, Les Abymes, Guadeloupe, France

Corresponding Author: avegarua@pasteur-guadeloupe.fr

Aedes aegypti immature stages develop in breeding sites where they ingest a wide variety of microorganisms, including bacteria. Here, we evaluated how the bacteria from breeding sites shape the microbiota of larvae and adults *Ae. aegypti*, as well as their impact on the transmission of some flaviviruses by this mosquito.

Impact of amoxicillin/clavulanic acid on *Aedes aegypti* microbiota and its capacity to transmit dengue virus

Margot GARCIA-VAN SMEVOORDE¹

¹ Institut Pasteur de Guadeloupe, Pointe-à-Pitre, Guadeloupe, France

Corresponding Author: mgarciavansmevoorde@pasteur-guadeloupe.fr

Dengue virus is the most common arbovirus worldwide especially in tropical and subtropical areas. Antibiotics are still widely used around the world. During a bite, *Aedes aegypti* can ingest blood containing antibiotics. What could be the impact on the microbiota and the transmission of dengue virus by the mosquito?

Disentangle chimeric (d-chimer) sequences in de novo assembled viromes

Sourakhata TIRERA¹

¹ Laboratoire des Interactions Virus-Hôtes, Institut Pasteur de la Guyane, Cayenne, French Guiana

Corresponding Author: stirera@pasteur-cayenne.fr

High-throughput sequencing, molecular biology and bioinformatics tools have contributed to the understanding of viral ecology. Nevertheless, amplification techniques and de novo assembly both introduce chimeras in the resulting contigs and bias the detected viral diversity particularly when “best-hit” (taxonomic assignment with the best BLAST score) is applied to BLAST outputs. Although methods were developed to deal with datasets that include chimeric sequences, few were adapted to viral shotgun metagenomics. Here, we propose a new tool, called “d-chimer”, which uses homology search via BLAST to perform taxonomic assignment and identify viruses in chimeric sequences. This approach has two main features: it enables more than one gene/organism to be kept per contig after homology search and it searches for uncovered zones recursively. Two de novo assembled virome datasets, from bird feces and rodent organs were analyzed with d-chimer and the commonly implemented “best-hit” approach. Compared to the “best-hit” approach, d-chimer succeeded in assigning 38.57% and 20.09% more viral fragments for bird and rodent raw outputs, respectively. Applying more stringent filtering on these BLAST hits, d-chimer still showed 19.34% and 12.46% more viral sequences, demonstrating a significant gain compared to the “best-hit” approach. The gain in the number of viral sequences obtained using d-chimer also impacts the viral diversity, with a significant increase in richness at the species (over 78% for bird and 50% for rodent datasets), genera (over 75% for bird and 46% for rodent datasets) and family (over 20% for bird dataset) levels. d-chimer proves to be efficient in the taxonomic assignment of shotgun viromes with chimeric sequences and may enable the exploration of rare viral sequences, and possibly, rare viral species.

Mini symposium 2: Multi-omics integration: challenges and perspectives

Organization

Silvia Bottini, Christophe Bécavin, Anaïs Baudot, Laura Cantini, Vincent Guillemot

Graph-based multi-omics integrative approaches to study microbial ecosystems across scales

Samuel CHAFFRON¹

¹ Lab Université de Nantes, Nantes, France

Corresponding Author: samuel.chaffron@univ-nantes.fr

Microorganisms form complex communities of interacting organisms sustaining key services across various ecosystems. Within us, human gut microbial communities are composed of essential core commensal bacteria that have a fundamental impact on our physiology, immune system, and health. In the ocean, marine plankton form the base of the food web, which sustain biogeochemical cycles, and help regulate climate. Understanding the mechanisms controlling microbiome assembly and sustaining their activities is a major challenge in microbial ecology. Though global meta-omics surveys are starting to reveal ecological drivers underlying microbiome community structures, it is unclear how community-scale species interactions are constrained, and how they are linked to ecosystem health. While multi-omics experimental approaches are becoming common practice, new integrative approaches need to be developed to enable the multi-scale characterisation of microbiomes. To go beyond statistical models, genome-resolved community networks enable to model and predict metabolic cross-feedings within prokaryotic assemblages. These mechanistic models allow to predict potential interactions within predicted communities and pinpoint specific metabolic cross-feedings shaping microbial communities. Integrating ecological and metabolic models provide a useful framework to assess community structure and organismal interactions, to reveal important mechanisms shaping natural microbial communities. The integrative analysis of heterogeneous multi-omics datasets generally allows to acquire additional insights and generate novel hypotheses about biological systems.

Multi-omic data integration with prior knowledge to decipher signaling and metabolic deregulation in complex diseases

Aurélien DUGOURD¹

¹ Saez Lab, Heidelberg University, Heidelberg, Germany

Corresponding Author: aurelien.dugourd@bioquant.uni-heidelberg.de

Microorganisms form complex communities of interacting organisms sustaining key services across various ecosystems. Within us, human gut microbial communities are composed of essential core commensal bacteria that have a fundamental impact on our physiology, immune system, and health. In the ocean, marine plankton form the base of the food web, which sustain biogeochemical cycles, and help regulate climate. Understanding the mechanisms controlling microbiome assembly and sustaining their activities is a major challenge in microbial ecology. Though global meta-omics surveys are starting to reveal ecological drivers underlying microbiome community structures, it is unclear how community-scale species interactions are constrained, and how they are linked to ecosystem health. While multi-omics experimental approaches are becoming common practice, new integrative approaches need to be developed to enable the multi-scale characterisation of microbiomes. To go beyond statistical models, genome-resolved community networks enable to model and predict metabolic cross-feedings within prokaryotic assemblages. These mechanistic models allow to predict potential interactions within predicted communities and pinpoint specific metabolic cross-feedings shaping microbial communities. Integrating ecological and metabolic models provide a useful framework to assess community structure and organismal interactions, to reveal important mechanisms shaping natural microbial communities. The integrative analysis of heterogeneous multi-omics datasets generally allows to acquire additional insights and generate novel hypotheses about biological systems.

Imaging-genetic approaches to study the human brain

Cathy PHILIPPE¹

¹ Neurospin, CEA, Saclay, France

Corresponding Author: cathy.philippe@cea.fr

Because of its critical part in life functions and its remarkable complexity, the brain is a sanctuary organ, which hinders its observation and study. Brain MRI offers a window into the living and functioning brain and provides various valuable intermediate phenotypes (endophenotypes) to study in relation to genomic data. To study the genetic roots of language in humans, we extracted various language-related endophenotypes and tested them for associations with SNPs in the largest imaging-genetic cohort in the general population, UKBiobank. In neuro-oncology, The Cancer Genome Atlas provides multi-omic data in Low-Grade Glioma (LGG), allowing us to analyze them jointly to predict survival better. We used the Sparse Canonical Correlation Analysis (SGCCA) framework, adding a graph constraint to get interpretable models regarding biological pathways. To study Alzheimer's disease, we leveraged the same methodology with the graph constraint applied to genotype data to relate neuroimaging features to genetic biomarkers and predict disease conversion.

Multi-omics data integration methods: kernel and other machine learning approaches.

Nathalie VIALANEIX¹

¹ Unité MIA-T (SaAB team and Bioinformatics platform), INRAe Toulouse, France

Corresponding Author: nathalie.vialaneix@inrae.fr

The substantial development of high-throughput biotechnologies has rendered large-scale multi-omics datasets increasingly available. New challenges have emerged to process and integrate this large volume of information, often obtained from widely heterogeneous sources. In this presentation, I will make a brief review of popular data integration methods and then focus on kernel methods and why they are usually well suited to this task.

Session 1: Structural bioinformatics and proteomics / Systems biology and metabolomic

The Martini Database (MAD): a web service to facilitate molecular simulations

Romuald Marin, Cecile Hilpert, P.C.T Souza, Luca Monticelli and Guillaume Launay

Institut de Biologie et Chimie des Protéines, 7 Passage du Vercors, 69007, Lyon France

Corresponding Author: romuald.marin@ibcp.fr

Abstract Coarse-grained (CG) models, such as the Martini force field, have emerged as a solution for reducing computational costs while capturing important aspects of molecular behavior. The Martini Database (MAD) (<https://mad.ibcp.fr/>) is a web service that simplifies the process of preparing CG simulations with Martini forcefield by providing verified and validated representations of various CG molecules, as well as tools for converting all-atom structures to CG representations and setting up large biological systems in a simulation box. Additionally, the new Polymer Editor tool helps scientists design new polymers or modify existing molecules. MAD's features can be accessed programmatically through an API, enabling automation and integration with other tools and systems.

Keywords website, database, structural bioinformatic, coarse-grained, molecular dynamics

1 Introduction

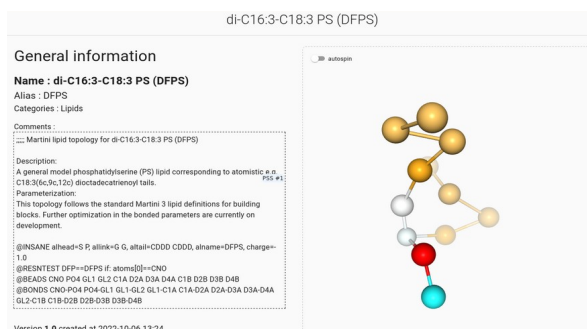
Molecular dynamics (MD) simulation is a powerful computational tool used to study the behavior of molecules over time, allowing us to see the impossible with conventional experimental methods. However, the computational cost of simulating forces and motions in every single atom can be prohibitive, particularly for large systems or long simulation times. To address this challenge, coarse-grained (CG) models have emerged as an attractive solution. One of the most popular CG models is the Martini force field [1], which uses a two to four-to-one mapping to group non-hydrogen atoms into larger entities called "beads," significantly reducing computational costs while still capturing important aspects of molecular behavior.

The Martini force field has undergone several updates over the years, and the community of users continues to grow. The Martini Database [2] (MAD) is a web service dedicated to sharing structures and topologies of molecules parameterized with the Martini CG force field. It also provides useful tools for submitting or retrieving CG representations of a wide range of molecules, transforming all-atom protein structures into CG structures and topologies, and assembling biomolecules into large systems. In the future, generating or editing polymers, and also modifying existing molecules, including polymeric post-translational modification in proteins, will be available in MADs. With these features, MAD simplifies the process of preparing CG simulations and enables users to dive into the Martini coarse-grained world with ease.

2 Overview of the website

2.1 Database content and access

This web service provides an open database designed to contain verified and validated representations of various CG molecules. Any registered user can submit a new CG model, which undergoes a curation process by Martini developers. Each entry in the database corresponds to a molecule and its corresponding CG files (topology and coordinate files) and force field version. Users can search and browse the database by force field, creation method, and biochemical category. Each molecule has a description page with general information, interactive molecular view, details, and version tracking.



di-C16:3-C18:3 PS (DFPS)

General information

Name : di-C16:3-C18:3 PS (DFPS)

Alias : DFPS

Categories : Lipids

Comments :

- Martini lipid topology for di-C16:3-C18:3 PS (DFPS)

Description:

A general model phosphatidylserine (PS) lipid corresponding to atomistic n , n , n PS#1.

C18:3([c,c,c,12c] dioctadecatrienoyl tails.

Parameterization:

This topology follows the standard Martini 3 lipid definitions for building blocks. Further optimization in the bonded parameters are currently on development.

@INSANE ahead+S P; allink+G G; altail+COO CD00; aName=DFPS; charge=1.0

@RESNTEST DFP==DFPS if: atoms[]==CNO

@BEADS CNO P04 GL1 GL2 C1A D2A D3A D4A C1B D2B D3B D4B

@BOND CNO P04 P04 GL1 GL1 GL2 GL1 C1A C1A D2A D2A D3A D3A D4A GL2 C1B C1B D2B D2B D3B D3B D4B

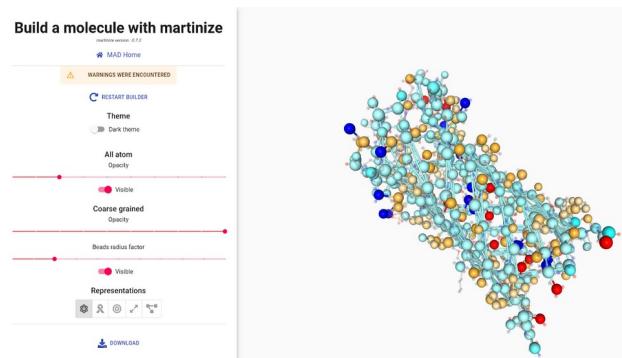
Version 1.0 created at 2022-10-06 13:24.

Screenshot of the molecule description section in MAD

An API is also available for programmatically accessing molecules in the database, allowing users to download different types of files for molecule representation and retrieve force field information. This feature enables automation and integration with other tools and systems. Third-party developers can use the data and functionality of the Martini Database to build new applications.

2.2 Molecule builder

Converting an all-atom representation to a CG representation is one of the most challenging problems for new users of the CG model. To address this issue, we designed a MAD feature based on the program martinize2 [3]. This feature allows users to generate CG representations (structures and topologies) from all-atom structures. The conversion process is fully automated, but advanced users can adjust a wide range of parameters through a control panel. The tool also provides a viewer for all-atom and CG representation and an option to edit output structures. The tool supports various versions of the Martini force field. It allows for the interactive editing of distance restraints and terminal protein modifications to improve protein stability and accuracy during molecular dynamics simulations. Once the CG molecule is generated, it is automatically saved to the user's history.



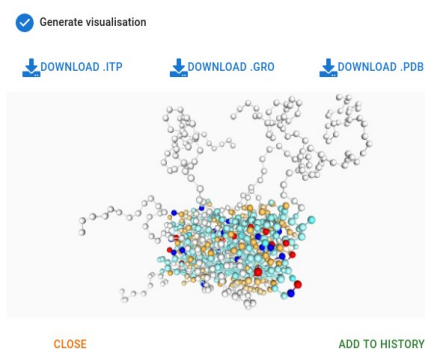
Screenshot of a CG Protein made with MAD: Molecule Builder

2.3 System Builder

This feature powered by Insane [4], allows users to set up large macromolecular systems in a simulation box. It enables combining different structures, such as models from the MAD Database, user-uploaded topology/structure files, and CG molecules from the user's private stash. It can produce different types of systems, such as phospholipid bilayers, any molecule present in MAD in a water solution or embedded in a lipid bilayer. The System Builder provides advanced controls and a view of the computed system with visualization options. The representation of the computed system can then be downloaded.

2.4 Future update: Polymer Editor

One challenge for molecular dynamics scientists is designing new polymers or modifying existing molecules, especially for those without strong computing skills. Our impending feature : Polymer Editor ; can help generate new polymer through a graphical interface, as opposed to using command line interfaces or complex file formats. The interface allows users to add small molecules and connect them to generate new polymers or import a molecule and link it to a polymer. The Polymer Editor tool is powered by Polyply [5] and provides a database of small biomolecules such as amino acids, sugars, and more, which can be used to create or modify molecules.



Screenshot of a PEGylated protein created with MAD:Polymer Editor

References

- [1] Siewert J. Marrink, 2023, WIREs Comput Mol, Two decades of Martini: Better beads, broader scope. <https://doi.org/10.1002/wcms.1620>
- [2] Hilpert Cecile, 2022, Facilitating CG simulations with MAD: the MArtini Database Server, BiorXiv. <http://dx.doi.org/10.1101/2022.08.03.502585>
- [3] KROON Peter, 2022, arXiv, Martinize2 and Vermouth: Unified Framework for Topology Generation, <https://doi.org/10.48550/arXiv.2212.01191>
- [4] Tsjerk A. Wassenaar, 2015, J. Chem. Theory Comput., Computational Lipidomics with insane: A Versatile Tool for Generating Custom Membranes for Molecular Simulations, <https://doi.org/10.1021/acs.jctc.5b00209>
- [5] Fabian Grünewald, 2022, Nat Commun, Polyply; a python suite for facilitating simulations of macromolecules and nanomaterials, <https://doi.org/10.1038/s41467-021-27627-4>

On the predictability of A-minor motifs from their local contexts

Coline GIANFROTTA^{1,2}, Vladimir REINHARZ³, Olivier LESPINET⁴, Dominique BARTH¹ and Alain DENISE^{2,4}

¹ Université de Versailles Saint-Quentin-en-Yvelines, Université Paris-Saclay, DAVID lab, Versailles, France

² Université Paris-Saclay, CNRS, Laboratoire Interdisciplinaire des Sciences du Numérique, Orsay, France

³ Department of Computer Science, Université du Québec à Montréal, Québec, Canada

⁴ Université Paris-Saclay, CNRS, I2BC, Orsay, France

Corresponding author: coline.gianfrotta@universite-paris-saclay.fr

Reference paper: Gianfrotta *et al.* (Nov. 2022) On the predictability of A-minor motifs from their local contexts, *RNA Biology*, 19:1, 1208-1227, <https://doi.org/10.1080/15476286.2022.2144611>

Introduction RNA molecules intervene in all major cellular processes. Their functions are strongly related to their three-dimensional structures. These structures are composed of a rigid skeleton, a set of canonical interactions called the secondary structure. On top of the skeleton, the nucleotides form an intricate network of interactions that are not captured by present thermodynamic models [1]. This network has been shown to be composed of modular motifs, that are linked to function, and have been leveraged for better prediction and design [2, 3]. A peculiar subclass of structural motifs are those connecting RNA regions far away in the secondary structure. They are crucial to predict since they determine the global shape of the molecule, therefore important for the function. Among them, the *A-minor motif* is particularly difficult to predict. Present in many families of non-coding RNAs, this motif has been shown to be important in the spatial folding of RNA molecules, as well as in cellular mechanisms such as codon-anticodon recognition during translation [4].

Purpose and method Our study focuses on the type I/II A-minor motif. It investigates the importance of the structural context in the formation of this motif. Several kinds of structural contexts are considered: the 3D substructure around a motif, called *3D context*, and the set of canonical and non canonical interactions in the context, called *topological context*. Our purpose is to determine what kind of information, contained in the structural context, can be useful to characterize and predict the presence and the position of type I/II A-minor motifs.

In a first step, we develop an automated method to classify A-minor motif occurrences according to their 3D context similarities, and then to characterize them. In a second step, we model the topological context of A-minor motif occurrences and of their classes by graphs, and use it to study the possibility of predicting the presence of A-minor motifs by assuming that the 3D context is not known, but we only know the topological context and possibly sequence information.

Results Our approach leads to two main results: Firstly, we uncover new subclasses of A-minor motif occurrences according to their local 3D similarities. The majority of classes are composed of homologous occurrences, but some of them are composed of non-homologous occurrences, which could suggest an evolutive convergence. Secondly, we show that, for some A-minor motifs, the topology combined with a sequence signal is sufficient to predict their presence and their position. In most other cases, these signals are not sufficient for predicting the A-minor motif, however we show that they are good signals for this purpose.

References

- [1] A. Lescoute and E. Westhof. The interaction networks of structured RNAs. *Nucleic acids research*, 34(22):6587–6604, 2006.
- [2] G. Chojnowski, T. Waleń, and J. M. Bujnicki. RNA Bricks—a database of RNA 3D motifs and their interactions. *Nucleic acids research*, 42(D1):D123–D131, 2014.
- [3] C. Oliver, V. Mallet, R. Sarrazin-Gendron, V. Reinharz, W. L. Hamilton, N. Moitessier, and J. Waldispühl. Augmented base pairing networks encode RNA-small molecule binding preferences. *Nucleic acids research*, 48(14):7690–7699, 2020.
- [4] A. Lescoute and E. Westhof. The A-minor motifs in the decoding recognition process. *Biochimie*, 88(8):993–999, August 2006.

An agnostic analysis of the human AlphaFold2 proteome using local protein conformations

Alexandre G. DE BREVERN¹

¹ DSIMB Bioinformatics team, INSERM UMR_S 1134, BIGR, Université Paris Cité and Université des Antilles and Université de la Réunion, F-75014 Paris, France

Corresponding Author: alexandre.debrevern@univ-paris-diderot.fr

Paper Reference: de Brevern (2023) An agnostic analysis of the human AlphaFold2 proteome using local protein conformations, *Biochimie*, 207:11-19. <https://doi.org/10.1016/j.biochi.2022.11.009>.

Databases contain millions of protein sequences, while the three-dimensional (3D) structures of proteins are much more difficult to obtain experimentally. For more than 30 years, different computational approaches have been implemented to propose 3D structural models of proteins from their amino acid sequence, i.e. homology / comparative modelling, threading, *ab initio*, *de novo* approaches, and meta-servers.

At the CASP13 competition (2018), DeepMind company presented its new Deep Learning approach named AlphaFold. It won the Free Modelling category, i.e. the prediction of novel protein folds, whereas template-Based Modelling category, i.e. protein folds already found in the Protein Data Bank, was won by Zhang's group. Two years later, AlphaFold version 2 [1] obtained particularly remarkable results at CASP14; some models were within the uncertainties of the experimental resolution, an impressive result. This improvement combined the delicate use of evolution, contacts within proteins, and large GPU computing power with a particularly complex and elegant architecture. AlphaFold 2 was since a hot topic, leading to a revolution in protein structural model building and opening new opportunities. Three points can be noticed (i) the code is on GitHub, (ii) different online notebooks are easy to use (e.g. CollabFold), and (iii) EBI provides structural model databases [2]. 98.5% of the human proteome is proposed with 36% the models are considered to be of atomistic quality, i.e. excellent for drug design and molecular dynamics.

The work carried out here does not consist of a new questioning of AlphaFold's result but provided a general view of its qualities. The human protein models provided by AlphaFold [2] were analysed using its confidence index (pLDDT score), with classic secondary structure and finer analysis of local protein conformation, e.g. γ -turns, β -turns and bends, β -turn types, PolyProline II, helix curvatures, β -bulges, and also a structural alphabet, namely Protein Blocks (PB) [3].

The results are, as expected, particularly coherent with our current knowledge of AlphaFold's result [4]. For instance, the large majority of α -helices are well predicted with high pLDDT scores. However, some points can be discussed which could potentially lead to improvements in the future: (i) PolyProline II (PPII) helices are too often encountered with a low confidence index. PPII represent 4-5% of all residues and are important in protein-protein interactions, it could so be an issue to be poorly approximated. (ii) In a very surprising way, while β -turns (turns of 4 residues) are well predicted, 55% of γ -turns (turns of 3 residues) have very low pLDDT scores. (iii) Even more strikingly, 94.8% of cis ω angles associated with low pLDDT scores, i.e. AlphaFold is clearly unable to propose proper cis ω angles. (iv) In an unexpected way, β -sheet occurrence is lower than expected, but the occurrence of PB *d* (it corresponds to the geometry of β -sheet core) is completely in accordance with the expected frequencies. Also, there are potentially β -sheets that were not founded until the end, which would explain this low frequency.

References

1. John Jumper *et al.* Highly accurate protein structure prediction with AlphaFold. *Nature*, 596 :583-589, 2021.
2. Kathryn Tunyasuvunakool *et al.* Highly accurate protein structure prediction for the human proteome. *Nature*. 596:590-596, 2021.
3. Alexandre G. de Brevern, Catherine Etchebest, Serge Hazout. Bayesian probabilistic approach for predicting backbone structures in terms of protein blocks. *Proteins*, 41:271-287, 2000.
4. Kristoffer T. Bæk, Kasper P. Kepp. Assessment of AlphaFold2 for Human Proteins via Residue Solvent Exposure. *J Chem Inf Model*, 62 :3391-3400, 2022.

An Agent-Based Model of Monocyte Differentiation into Tumor-Associated Macrophages in Chronic Lymphocytic Leukemia

Nina VERSTRAETE^{1,2}, Malvina MARKU^{1,2}, Marcin DOMAGALA^{1,2}, H el ene ARDUIN^{1,2}, Julie BORDENAVE^{1,2}, Jean-Jacques FOURNI E^{1,2}, Loic YSEBAERT^{1,2,4}, Mary POUPOT^{1,2}, and Vera PANCALDI^{1,2,3}

¹ INSERM, CRCT, 2 Avenue Hubert Curien, 31037, CEDEX 1 Toulouse, France

² Universit e Toulouse-III Paul Sabatier, Route de Narbonne, 31330 Toulouse, France

³ Barcelona Supercomputing Center, Carrer de Jordi Girona, 29, 31, 08034 Barcelona, Spain

⁴ Service d'H ematologie, IUCT-Oncopole, 31330 Toulouse, France

Corresponding Author: nina.verstraete@inserm.fr

Verstraete et al. (2023) An Agent-Based Model of Monocyte Differentiation into Tumor-Associated Macrophages in Chronic Lymphocytic Leukemia, iScience 2023. <https://doi.org/10.1101/2021.12.17.473137>

In the tumor microenvironment, tumor-associated macrophages are known to play a critical role in the survival and chemoresistance of cancer cells. In the case of chronic lymphocytic leukemia (CLL), these tumor-associated macrophages are called Nurse-Like Cells (NLCs) and reside mainly in the lymph nodes, where they are able to protect leukemic B cells (B-CLL) from spontaneous apoptosis and contribute to their chemoresistance, hindering the efficacy of immunotherapy in many patients. NLCs are differentiated from monocytes through cytokines signaling and physical contact with the cancer cells [1], however, the precise mechanisms by which B-CLL cells influence this differentiation are still unknown. We used an *in vitro* model of leukemia, in which we can closely follow the production of NLCs from monocytes in the presence of leukemic B cells from CLL patients. Building on experimental observations of cancer cells in these cultures of patients' blood, we propose here a two-dimensional agent-based model simulating the monocyte-to-macrophage differentiation and inter-cellular interactions in the spatial context of this *in vitro* co-culture of monocytes and cancer B-CLL cells.

Using our time-course measurements of B-CLL cell viability and concentration to optimize the model parameters, we were able to reproduce the experimentally observed average dynamics. However, given the extreme variability between patients, we also opted for patient-specific parameter optimizations, identifying 2 distinct patient classes, which might correspond to protective and non-protective NLCs. We show that this unsupervised classification is consistent with experimental evidence, displaying a majority of M2 markers in the case of the protective NLCs, and a majority of M1 markers in the case of the non-protective NLCs [2]. The level of protective anti-apoptotic signals, that are known to be provided by NLCs to the cancer cells [1], appear to also be important to differentiate between protective and non-protective NLCs, in agreement with the fact that protective NLCs secrete more anti-apoptotic signals than non-protective ones. Finally, parameter sensitivity analysis and differences in parameter distributions between the 2 patient classes suggest a fundamental role of phagocytosis efficiencies from macrophages and NLCs, their sensing distances of dead and apoptotic cells and the movement probability of apoptotic cells to ensure the long-term survival of cancer cells in this *in vitro* CLL model. These findings suggest that monitoring and potentially modulating phagocytosis could play a role in the control of TAM formation *in vitro*, in CLL lymph nodes or even in solid tumors [3].

References

1. Fr ed eric Boissard et al. Nurse like cells: Chronic lymphocytic leukemia associated macrophages. *Leuk. Lymphoma*, 56(5):1570- 1572, 2015.
2. Marcin Domagala et al. Il-10 rescues cll survival through repolarization of inflammatory nurse-like cells. *Cancers*, 14(1):16, 2022.
3. Nina Verstraete et al. An Agent-Based Model of Monocyte Differentiation into Tumor-Associated Macrophages in Chronic Lymphocytic Leukemia. *iScience*, 2023. <https://doi.org/10.1101/2021.12.17.473137>

Multi-omics characterization of Richter syndrome unlocks classifiers and predictors of outcome into broader and heterogeneous lymphoma datasets

Sébastien HERGALANT¹, Romain PIUCCO¹, Ghislain FIÉVET¹, Emil CHTEINBERG², Stephan STILGENBAUER³,
Reiner SIEBERT², David MEYRE¹, Pierre FEUGIER¹ and Julien BROSÉUS¹

¹ Nutrition, Genetics and Environmental Risk Exposure (NGERE) – Inserm U1256, Campus Biologie Santé, Nancy, France

² Institute of Human Genetics, Ulm University Medical Center, Ulm, Germany

³ Department of Internal Medicine III, Ulm University, Ulm, Germany

Corresponding Author: sebastien.hergalant@inserm.fr

Reference paper: Broséus & Hergalant *et al.* (2023) Molecular characterization of Richter syndrome identifies *de novo* diffuse large B-cell lymphomas with poor prognosis, *Nature Communications*, 2023, 14, 309.
<https://doi.org/10.1038/s41467-022-34642-6>

Richter syndrome (RS) is the onset of a dismal diffuse large B-cell lymphoma (DLBCL) subtype that exemplifies aggressiveness and chemoresistance occurring in the context of indolent chronic lymphocytic leukemia (CLL). In this study we characterize a large series of human RS samples [1] by genome-wide DNA methylation and whole-transcriptome profiling from a multi-omics setup including additional copy number alterations, exome and proteome data. We comprehensively compare them to i) the paired CLL component, i.e. samples acquired before the aggressive transformation in the same patients, ii) a broad CLL reference group [2] and iii) *de novo* DLBCLs of different cell-of-origin (COO) and molecular classification.

Recent genomic studies combining DNA and RNA sequencing extended DLBCL subtyping beyond COO, identifying subgroups defined by their genomic alteration patterns and associated clinical courses, but a notable proportion remains unclassified [3]. Distinguishing between CLL-derived RS and *de novo* DLBCL in a diagnostic setting based on histology and immunochemistry alone is challenging. Most RS cases arise from the preceding CLL clone, while the remainder are in fact independent *de novo* DLBCLs, a dichotomy of importance for treatment decisions. Indeed, *de novo* DLBCLs are chemosensitive in most patients, whereas CLL-derived RS are chemoresistant, with a median overall survival of around 12 months.

Adjusting for the lack of appropriate human or animal models to study RS, our integrative approach provides insights into its epigenomic architecture, corroborates two evolutionary groups of RS [4] and unravels a CLL epigenetic imprint in clonally related samples. Thus removing the need for the initial CLL tumor DNA, a significant improvement since CLL stage is often undiagnosed and/or corresponding samples unavailable. We also define two novel classifiers: a methylation-based predictor to detect a stable CLL “memory” over disease evolution, and a gene-expression-based scoring method outlining a novel DLBCL subgroup from public datasets harboring this CLL-derived RS epigenetic imprint. Applying both classifiers to omics data from landmark studies uncovers a subset of “RS-type” DLBCL enriched in cases with a specific COO signature, unclassified or undetected by other genomic classifiers, and with the same unfavorable prognosis as RS. These findings directly translate prognostication of *de novo* DLBCLs, the most common human B-cell lymphoma, and associate them linearly with overall and progression-free survival, independently of known clinical factors and biological covariates.

References

1. Charline Moulin, Francis Guillemin, Thomas Remen, Florian Bouclet *et al.* Clinical, biological, and molecular genetic features of Richter syndrome and prognostic significance: A study of the French Innovative Leukemia Organization. *American Journal of Hematology*, 96(9), 311-314, 2021.
2. Renée Beekman, Vincente Chapaprieta, Nùria Russiñol, Roser Vilarrasa-Blasi *et al.* The reference epigenome and regulatory chromatin landscape of chronic lymphocytic leukemia. *Nature medicine*, 24(6), 868-880, 2018.
3. George Wright, Da Wei Huang, James Phelan, Zana Coulibaly *et al.* A probabilistic classification tool for genetic subtypes of diffuse large B cell lymphoma with therapeutic implications. *Cancer cell*, 37(4), 551-568, 2020.
4. Erin Parry, Ignaty Leshchiner, Romain Guièze, Connor Johnson *et al.* Evolutionary history of transformation from chronic lymphocytic leukemia to Richter syndrome. *Nature medicine*, 29, 158–169, 2023.

Session 2: Evolution

Horizontal transfer shapes the subsequent emergence of antibiotic resistance by point mutations

Charles COLUZZI¹, Martin GUILLEMET¹, Maxime GODFROID², Guillaume ACHAZ², Philippe GLASER³, Eduardo P.C. ROCHA¹

¹ Institut Pasteur, Université de Paris Cité, CNRS, UMR3525, Microbial Evolutionary Genomics, Paris, 75015, France.

² Center for Interdisciplinary Research in Biology (CIRB), Collège de France, CNRS, INSERM, Université PSL, SMILE Group, Paris, 75005, France.

³ Institut Pasteur, Université de Paris Cité, CNRS, UMR6047, Unité EERA, Paris, 75015, France.

Corresponding Author: charles.coluzzi@pasteur.fr

Abstract *Bacterial populations can adapt quickly to environmental challenges by genome modifications. These can be endogenous, such as mutations, or exogenous, such as horizontal gene transfer. Antibiotic resistance is a perfect example of this process, as it can arise in both ways. However, epidemiological data reveal that certain bacterial lineages are more susceptible to acquiring antibiotic resistance via point mutations than others. This suggests that adaptation, even in core genes, requires appropriate genetic backgrounds. Yet, bacterial genetic backgrounds change rapidly due to horizontal gene transfer and gene loss, and it is challenging to ascertain the evolutionary processes leading to antibiotic resistance. In this study, we used a likelihood framework to infer the correlated evolutionary events leading to antibiotic resistance in 1600 *Escherichia coli* genomes. We inferred the chronologies of past events to test if specific genes were consistently acquired through horizontal transfer before the acquisition of point mutations in core genes that led to quinolone resistance. We found numerous distinct groups of genes in genetic linkage that were statistically more frequently acquired or lost prior to the acquisition of quinolone resistance. These findings suggest that bacterial lineages followed different but parallel evolutionary paths that led to antibiotic resistance. Hence, horizontal gene transfer enables the emergence of genetic backgrounds that favor subsequent phenotypic evolution by point mutations in core genes.*

Keywords antibiotics resistance, evolutionary trajectories, epistasis interactions, horizontal gene transfer, genetic background

1. Introduction

Bacterial populations adapt rapidly to challenges such as bacteriophage predation, antibiotic therapy, or environmental perturbations. This process is facilitated by endogenous genome modifications such as mutations or deletions and by genetic exchanges with other bacteria. Since the genome is made of thousands of genes linked by complex regulatory interactions and encoding proteins with multiple physical interactions, these modifications may be costly. The trade-off between the benefits and the costs of the novel variants determines the outcome of natural selection of novel variants. Many studies have started to unravel how the acquisition of novel functions by horizontal gene transfer (HGT) depends on the existing genetic background [1, 2]. The role of epistatic interactions between the genetic background and genes acquired horizontally in shaping the probability of fixation of transfers is important because bacterial gene repertoires vary rapidly by HGT. For example, less than half of the average *Escherichia coli* genome corresponds to genes present in more than 99% of the strains (core genome). The large genomic variation caused by HGT means that genetic backgrounds can be very different within a species. Hence, HGT may be affected by the genetic background, but the evolutionary trajectories of conserved genes such as those of the core genome may also be affected by previous changes in gene repertoires by HGT. The existence of the latter epistatic interactions has been little studied.

Antibiotic resistance in human pathogens (and commensals) is an example of the ability of bacteria to adapt to novel challenges by a mixture of HGT and mutational processes. Bacteria can modify the target of the

antibiotic by mutation, inactivate the antibiotic by acquisition of appropriate enzymes by HGT, or diminish its cellular concentration using efflux pumps. These mechanisms may pre-exist in the genome or be acquired by mutation or HGT [3]. Epistatic interactions at the level of single genes were shown to shape the emergence of antibiotic resistance [4]. They constrain the acquisition of resistance by favoring certain mutational paths that result in intermediate steps with lower-than-average fitness costs and higher than average resistance. Much less is known on how antibiotic resistance is shaped by the bacterial genetic background. In a seminal study, the fitness cost of chromosomal resistance to several antibiotics acquired by point mutations was in negative epistatic association with the presence of conjugative plasmids in more than half of the tested combinations [5]. Hence, interactions between genes acquired by HGT and chromosomal mutations may be important. Given the differences in terms of genetic backgrounds across a bacterial species, caused by rampant HGT, epistatic interactions may contribute to explain why certain lineages are often much more likely to be antibiotic resistant than others [6, 7].

Here, we wished to know if the changes in the genetic background potentiated by HGT could have an impact on the fixation of subsequent mutations in the core genes that result in antibiotic resistance. We focused on the mutations conferring resistance to quinolones because these are well known and occur in core essential genes. Quinolones stabilize the breaks made in the DNA by the gyrase or topoisomerase, and the resulting complex inhibits DNA synthesis [8]. The mechanisms that provide highest resistance to quinolones have been characterized in detail and involve mutations in target proteins. We searched for these mutations in a large datasets of *E. coli* genomes that consists of 1268 complete *E. coli* from NCBI RefSeq and it is usually presumed that it contains many clinical strains. This allowed to assess the prevalence of these mutations and the order of their acquisition. Having this information, we used phylogenetic methods to identify if there are genes acquired by HGT at earlier stages that influence the acquisition and fixation of quinolone mutations in the species. The goal of our work was not just to identify genome-wide associations between the mutations and other genes, we aimed to identify genes acquired unmistakably before the resistance mutations. This way, we can establish chronologies between the acquisition of genes and the acquisition of antibiotic resistance. Studying these chronologies will improve our global understanding of the evolution of antibiotic resistance and allow us to identify the emerging threats of antibiotic resistance.

2. Materials & Methods

2.1 Genome data, pangenome, core-genome and phylogeny

We analyzed 21 086 complete genomes retrieved from NCBI RefSeq representing 6 124 species of Bacteria (<http://ftp.ncbi.nih.gov/genomes/refseq/bacteria/>), in March 2021. From these bacteria, we retrieved 1 585 genomes that were identified as *E. coli*. The genomes were analyzed for assembly quality using their L90 value and for genetic distance to using MASH [9]. We removed from further analysis strains because they had a L90 superior to 100 or because they were too divergent (MASH distance > 6%) or too similar (MASH distance < 0.0001) to other strains. This resulted in a dataset of 1 268 completely sequenced and assembled *E. coli* genomes. Furthermore, the pangenomes and core-genomes of the 1 268 *E. coli* were computed using PanACoTA v.1.3.1 [10]. Briefly, the pangenome was constructed using MMseqs2 (protein identity >80%) and consist of 78653 families. From this pangenome, the core-genome was retrieved. It consists of 2393 gene families present in exactly 1 copy in at least 99.0% of all genomes. In the remaining 1% genomes, there can be 0, 1 or several members of the gene family.

Phylogeny was reconstructed based on the concatenated DNA alignments of the core genes. The phylogenetic analyses were performed using IQ-TREE [11] with the ultra-fast bootstrap option (-bb 1000 bootstraps) and with the best fitting model estimated using *ModelFinder Plus* (-MPF). The best model was GTR+F+I+G4 according to the BIC criterion. Trees were rooted using the midpoint function from the phangorn packages (v.2.5.5) for R.

2.3 Detection of quinolone resistance conferring mutations

Point mutations leading to quinolone resistance in *E. coli* are found in the DNA gyrase and the DNA topoisomerase IV genes, i.e., in *gyrA*, *gyrB*, *parC* and *parE* [8]. We retrieved the sequences of these gene from the proteome of each dataset using blastp and aligned them using the *MAFFT* (with -auto parameters). The alignments were parsed using Biopython [12]. We looked for point mutations leading to quinolone resistance,

at each expected position (accounting for gaps) using the comprehensive list provided by Hopkins et al. [8] for *E. coli*.

2.4 Inference of ancestral gene repertoires

We counted the number of occurrences of each family of the pan-genome in all genomes. This was used to build a gene content occurrence matrix in all the taxa (leaves of the phylogenetic trees). From this occurrence matrix, we inferred the ancestral state (presence or absence) of each gene family at every internal node of the phylogenetic trees with PastML v1.9.33 [13]. We used the JOINT method with default parameters, this method reconstructs the states of the scenario with the highest likelihood. From the ancestral state matrix computed by PastML, we inferred the gene gains and losses for all the branches of the tree by subtracting the gene content of the child node to the gene content of the parent node.

2.5 Trajectories of acquisition of mutations

We build the history of the quinolone resistance mutations acquisition. In this case, we built a presence/absence matrix of each type of mutation in each genome. Ancestral states for each mutation were inferred the same way as described in the previous section. To observe the distinct chronologies of the mutation acquisition, every path leading to a leaf containing at least one quinolone resistance mutation was extracted using the ape package v5.3 for R. Paths were then traversed from the root to the first acquisition of mutation conferring quinolone resistance (as inferred by pastML). This was considered as the first mutation acquired on this path. All paths arising from this first mutation were then traversed from this event of acquisition to the next event on the path and this recursively until the last event of acquisition on each path. This way, we only consider the number of events in the tree, not the number of taxa affected by them. Hence, for a given series of changes, it will produce just as much signal whether it is at the root of a large clonal lineage or leading to a single genome. Therefore, it was not necessary to trim the tree of clonal lineages, as they do not bias the results

This resulted in a collection of paths corresponding to every single sequence of mutation acquisition. We summed every identical chronology and summarized them in a graph.

2.6 Detection of epistatic interactions

Epistasis in the acquisition of quinolone resistance was investigated by testing for independence between the events of acquisition of this resistance and the gain or loss of each gene family of the pangenome (see section above). We carried out this analysis using the program Evo-Scope [14] which compares the number of occurrences of an event E_1 (e.g. acquisition of resistance) following an event E_0 (e.g. gain of a certain gene) on a tree to the expected number under a null model of uniform rates on the tree [15]. Once we identified joint evolution between pairs of events, we inferred the type and strength of correlated evolution between them. The interactions were classified according to 3 different scenarios to determine which of the trait influenced the occurrence of the other one.

-Scenario of independence: The occurrence of event E_0 does not change the occurrence rate of E_1 .

-Scenario of asymmetric induction: The occurrence of event E_0 increase the occurrence rate of event E_1 .

-Scenario of reciprocal induction: Event E_0 enhance the occurrence rate of event E_1 and, reciprocally, event E_1 enhance the rate of occurrence of event E_0 .

The multiple scenarios can contain from two to eight parameters and describe the interrelationship between the occurrence of two events on the tree. The parameters are divided into natural occurrence rates and excited occurrence rates for each of the trait. The ratio λ between the excited occurrence rates and the natural occurrence rates defines the induction. If $\lambda > 1$, the induction is positive (i.e., the occurrence of the event E_1 is increased after the event E_0) whereas it is negative when $\lambda < 1$.

2.7 Constructing groups of genes in genetic linkage

Pangenome families consistently acquired before the resistance to quinolone were clustered. First, pangenome families were clustered in groups if they were consistently co-acquired or co-lost in the same branch or node of the tree. This was assessed using the Epics module of the program Evo-Scope with the parameter -I. Thus, the program compares the number of occurrences of an event E_1 at the same branch or

node as the occurrence of an event E_0 to the expected number under a null model of uniform rates on the tree. This way, we identified pairs of events consistently occurring together. Pairs of events identified that way were then clustered by single-linkage using the agglomerative clustering algorithm from scikit-learn v1.2.2 (parameters: affinity='precomputed', linkage='single').

The precedent groups were then split in relation to their localization in the genome. For every pair of gene families in a cluster of co-acquisition, we computed the median distance between the genes in the genome (when they co-occurred). These values were then used to cluster the gene families by average-linkage using the agglomerative clustering algorithm from scikit-learn v1.2.2 (parameters: affinity='precomputed', distance_threshold=30, linkage='average', n_clusters=None).

3. Results

3.1 Frequency of the mutations providing resistance to quinolones

We extracted from the proteome of the 1268 *E. coli* the sequences of the *gyrA*, *gyrB*, *parC* and *parE* genes, which are also called the quinolone-resistance determining regions (QRDR). We screened these regions for 39 different quinolone resistance conferring mutations obtained from the literature [8].

Mutations had very different prevalence with some that were remarkably frequent. The most common mutations were found in the *gyrA* and *parC* genes. For instance, the *gyrA*_S83L mutation which confers high resistance at a significant fitness cost is present in 34% of the strains. The two other most frequent mutations, the *parC*_S80I and the *gyrA*_D87N mutations were both present in more than 26% and 24% of the strains, respectively.

3.2 Chronologies of acquisition of the mutations

The phylogenetic analysis shows a strong co-occurrence of the different mutations conferring quinolone resistance with the combination of the *gyrA*_S83L-*parC*_S80I-*gyrA*_D87N being the most frequent. The combination of these three mutations is known to confer a high level of resistance with a limited fitness cost [16, 17], and it is widely distributed in our dataset (23.4% of the strains). Indeed, these three mutations co-occurred much more frequently together than separately.

If mutations typically accumulate and some combination are more frequent than others, one might expect that epistatic processes are at play. It is then possible that the accumulation of substitutions providing resistance follow typical chronologies. We inferred the chronology of acquisition of the five most common mutations in our dataset on the phylogenetic tree of the specie. In this analysis, we did not consider the reverse mutations for the sake of clarity (they are very rare). We identified 41 occurrences of lineages acquiring only the *gyrA*S83L mutation. Some of these mutations then gave rise to the triple resistant strain. The other most frequent trajectory was the one going from the fully sensitive combination to the triple resistant (29 occurrences). The third most common combination of mutations (*gyrA*S83L, *parC*S80I, *gyrA*D87N, *parE*I529L and *parE*C84V) only occurred once from a step-wise event of acquisition of *gyrA*S83L and *parE*I529L, followed by the acquisition of the other three mutations (Figure 1). Altogether, this suggests that the mutation *gyrA*S83L is the first one fixed in most lineages. It also suggests that all remaining mutations are then acquired very quickly, such that they are inferred to occur jointly in a single branch of the phylogenetic tree.

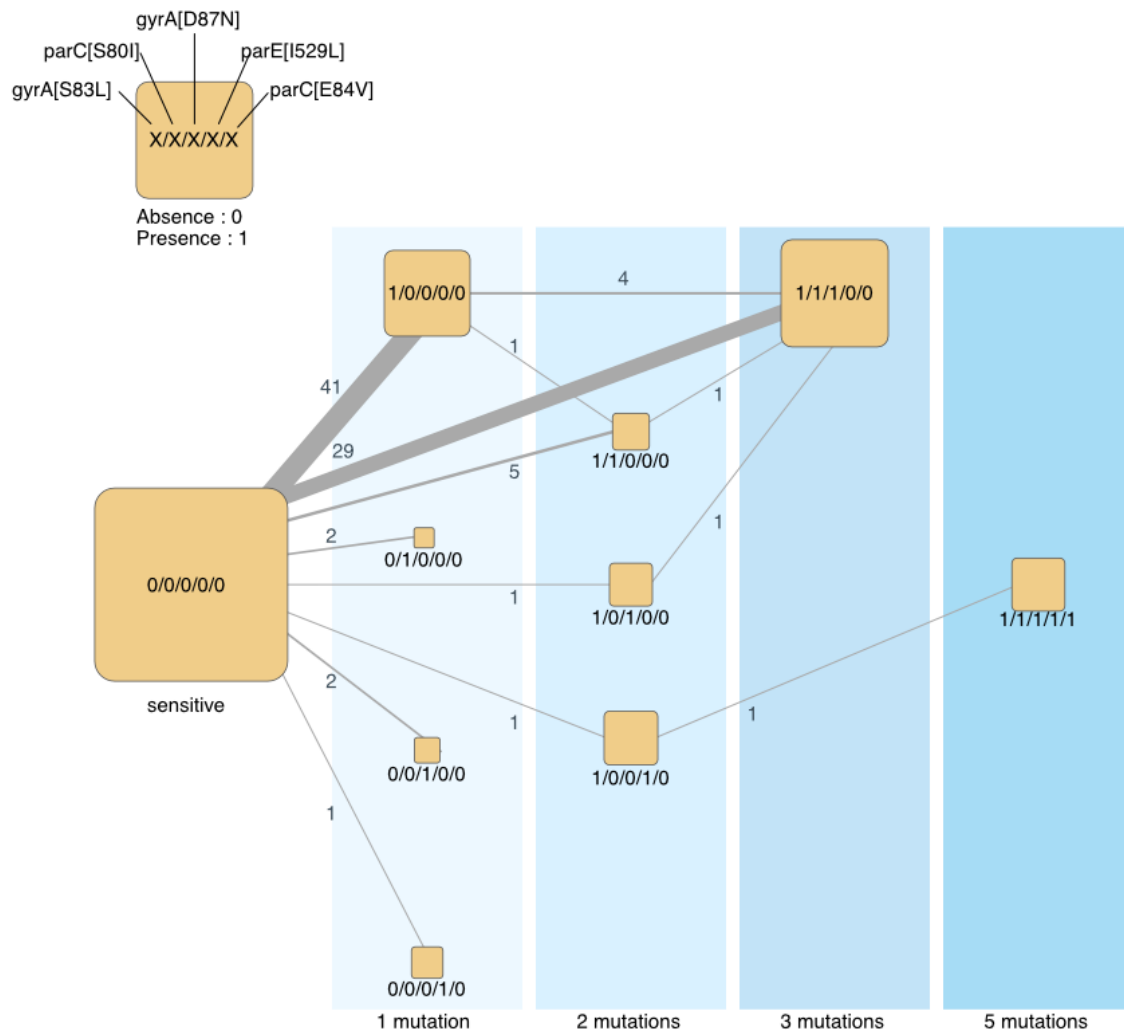


Figure 1: Trajectories of acquisition of the main resistance mutations in *E. coli*. The blue areas represent the number of distinct mutations in genomes. The size of the square scales with the number of genomes observed (leaves of the tree). The edges represent the chronologies of acquisition of one or several mutations as inferred from the reconstruction of ancestral states on the tree. The edge size is proportional to the frequency of the respective transition, with the labels showing this exact number.

3.3 Clusters of genes gained and lost before the acquisition of the resistance

To enquire on how the dynamics of gene repertoires may favor the acquisition of resistance to quinolones by point mutations in the core genes targeted by the antibiotics, we searched for genes that were frequently gained or lost before the emergence of these mutations. Since most mutations co-occurred, we defined taxa as resistant when they had at least one resistance mutation. We then reconstructed the ancestral states for the resistance, i.e. we inferred the branch where the resistance was acquired. We used a similar procedure to identify the branches in the tree where each family of the pangenome was lost or gained. We then used EvoScop [14] to compare the chronology of gains and losses of every family in the pan-genome with the acquisition of the resistance to the antibiotic and identify the significant chronologies, i.e. when one event is consistently followed by another event (in this case the second being the acquisition of resistance). This method identified 183 gene gains and 26 gene losses occurring consistently before the acquisition of the resistance ($p < 10^{-5}$ after correction for multiple tests).

(2) Influence on the acquisition rate of the resistance is assessed by the lambda values given by Evoscop (see methods). These values were not used for the clustering but are expected to be similar among genes under strong genetic linkage. Indeed, genes in clusters had homogeneous lambda values (Levene Test: $5.71e-06$, MSE (mean square of the error) = 0.591, one-way ANOVA ***), meaning that genes within the same cluster have a similar impact on the acquisition rate of the quinolone resistance. Out of the 60 clusters, the majority (49) promote the acquisition of the resistance, while only 11 clusters reduce it (Figure 2, bottom right)

(3) Genes under a strong genetic linkage are expected to be more likely gained or lost together. All the clusters were exclusively made of loss or gain of genes. Out of the 60 clusters, the majority (41) correspond to gene gains and 19 correspond to gene loss. Bacteria need to acquire all the genes involved in a metabolism to be able to express it. However, the loss of only few genes is sufficient to disrupt the activity of entire pathways. In our analysis, groups corresponding to gains of genes tend to be bigger than the ones corresponding to losses (Mann–Whitney Test: p -value = 0.0108) (Figure 2, bottom left)

4. Conclusion

In this work, we inferred the chronology of acquisition of the mutation conferring resistance to quinolone. We found that mutations were acquired in a preferential order suggesting the existence of epistatic interactions between the mutations. We then searched for events of gain or loss of genes prior the acquisition of point mutation conferring resistance to the quinolone. We found numerous groups of genes in genetic linkage that were consistently acquired or lost prior to the acquisition of the quinolone resistance. Additionally, these groups were phylogenetically distinct, yet they encoded recurrent biological functions and were encoded by different mobile genetic elements. Taken together, these findings suggest that these bacterial lineages followed different but parallel evolutionary paths that led to antibiotic resistance, which was favored by the pre-acquisition of certain genes by HGT.

Acknowledgements

This project was funded by the INCEPTION project (PIA/ANR-16-CONV-0005), Equipe FRM (EQU201903007835), Laboratoire d'Excellence IBEID (ANR-10-LABX-62-IBEID). This work used the computational and storage services (MAESTRO cluster) provided by the IT department at Institut Pasteur, Paris.



References

1. Press, M.O., et al., *Genome-scale co-evolutionary inference identifies functions and clients of bacterial Hsp90*. PLoS Genet, 2013. **9**(7): p. e1003631.
2. Szappanos, B., et al., *Adaptive evolution of complex innovations through stepwise metabolic niche expansion*. Nat Commun, 2016. **7**: p. 11607.
3. Davies, K.M. and P.J. Lewis, *Localization of rRNA Synthesis in Bacillus subtilis: Characterization of Loci Involved in Transcription Focus Formation*. J Bacteriol, 2003. **185**: p. 2346-53.
4. Weinreich, D.M., et al., *Darwinian evolution can follow only very few mutational paths to fitter proteins*. Science, 2006. **312**: p. 111-4.
5. Silva, R.F., et al., *Pervasive sign epistasis between conjugative plasmids and drug-resistance chromosomal mutations*. PLoS Genet, 2011. **7**(7): p. e1002181.
6. Leavis, H.L., M.J. Bonten, and R.J. Willems, *Identification of high-risk enterococcal clonal complexes: global dispersion and antibiotic resistance*. Current opinion in microbiology, 2006. **9**(5): p. 454-460.
7. Wyres, K.L., M.M. Lam, and K.E. Holt, *Population genomics of Klebsiella pneumoniae*. Nature Reviews Microbiology, 2020: p. 1-16.
8. Hopkins, K.L., R.H. Davies, and E.J. Threlfall, *Mechanisms of quinolone resistance in Escherichia coli and Salmonella: recent developments*. Int J Antimicrob Agents, 2005. **25**(5): p. 358-73.
9. Ondov, B.D., et al., *Mash: fast genome and metagenome distance estimation using MinHash*. Genome Biol, 2016. **17**(1): p. 132.
10. Perrin, A. and E.P.C. Rocha, *PanACoTA: a modular tool for massive microbial comparative genomics*. NAR Genom Bioinform, 2021. **3**(1): p. lqaa106.
11. Nguyen, L.T., et al., *IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies*. Mol Biol Evol, 2015. **32**(1): p. 268-74.

12. Cock, P.J. and D.E. Whitworth, *Evolution of relative reading frame bias in unidirectional prokaryotic gene overlaps*. Mol Biol Evol, 2010. **27**(4): p. 753-6.
13. Ishikawa, S.A., et al., *A Fast Likelihood Method to Reconstruct and Visualize Ancestral Scenarios*. Mol Biol Evol, 2019. **36**(9): p. 2069-2085.
14. Godfroid, M., et al., 2022.
15. Behdenna, A., et al., *Testing for Independence between Evolutionary Processes*. Syst Biol, 2016. **65**: p. 812-23.
16. Bagel, S., et al., *Impact of gyrA and parCMutations on Quinolone Resistance, Doubling Time, and Supercoiling Degree of Escherichia coli*. Antimicrobial agents and chemotherapy, 1999. **43**(4): p. 868-875.
17. Marcusson, L.L., N. Frimodt-Moller, and D. Hughes, *Interplay in the selection of fluoroquinolone resistance and bacterial fitness*. PLoS Pathog, 2009. **5**(8): p. e1000541.



Limited Transmission of *Klebsiella pneumoniae* among Humans, Animals, and the Environment in a Caribbean Island, Guadeloupe (French West Indies)

Alexis Dereeper,^a Gaëlle Gruel,^a Matthieu Pot,^a David Couvin,^a Elodie Barbier,^b Sylvaine Bastian,^c Jean-Christophe Bambou,^d Moana Gelu-Simeon,^e Séverine Ferdinand,^a Stéphanie Guyomard-Rabenirina,^a Virginie Passet,^f Frederic Martino,^g Pascal Piveteau,^h Yann Reynaud,^a Carla Rodrigues,^f Pierre-Marie Roger,^{ij} Xavier Roy,^k Antoine Talarmin,^a Benoit Tressieres,^l Marc Valette,^g  Sylvain Brisse,^f  Sébastien Breurec^{a,c,j,l}

^aTransmission, Reservoir and Diversity of Pathogens Unit, Pasteur Institute of Guadeloupe, Pointe-à-Pitre, France

^bUMR AgroEcologie, INRAE, Bourgogne Franche-Comté University, Dijon, France

^cLaboratory of Clinical Microbiology, University Hospital Center of Guadeloupe, Pointe-à-Pitre/Les Abymes, France

^dINRAE, ASSET, Petit-Bourg, France

^eHepato-Gastroenterology Department, University Hospital Center of Guadeloupe, Pointe-à-Pitre/Les Abymes, France

^fInstitut Pasteur, University Paris Cité, Biodiversity and Epidemiology of Bacterial Pathogens, Paris, France

^gIntensive Care Department, University Hospital Center of Guadeloupe, Pointe-à-Pitre/Les Abymes, France

^hUR OPAALE, INRAE, Rennes, France

ⁱInfectious Disease Department, University Hospital Center of Guadeloupe, Pointe-à-Pitre/Les Abymes, France

^jFaculty of Medicine Hyacinthe Bastaraud, University of the Antilles, Pointe-à-Pitre, France

^kVeterinary Clinic, Baie-Mahault, Guadeloupe

^lINSERM Center for Clinical Investigation 1424, Pointe-à-Pitre/Les Abymes, France

ABSTRACT Guadeloupe (French West Indies), a Caribbean island, is an ideal place to study the reservoirs of the *Klebsiella pneumoniae* species complex (KpSC) and identify the routes of transmission between human and nonhuman sources due to its insularity, small population size, and small area. Here, we report an analysis of 590 biological samples, 546 KpSC isolates, and 331 genome sequences collected between January 2018 and May 2019. The KpSC appears to be common whatever the source. Extended-spectrum- β -lactamase (ESBL)-producing isolates (21.4%) belonged to *K. pneumoniae sensu stricto* (phylogroup Kp1), and all but one were recovered from the hospital setting. The distribution of species and phylogroups across the different niches was clearly nonrandom, with a distinct separation of Kp1 and *Klebsiella variicola* (Kp3). The most frequent sequence types (STs) (≥ 5 isolates) were previously recognized as high-risk multidrug-resistant (MDR) clones, namely, ST17, ST307, ST11, ST147, ST152, and ST45. Only 8 out of the 63 STs (12.7%) associated with human isolates were also found in nonhuman sources. A total of 22 KpSC isolates were defined as hypervirulent: 15 associated with human infections (9.8% of all human isolates), 4 (8.9%) associated with dogs, and 3 (15%) associated with pigs. Most of the human isolates (33.3%) belonged to the globally successful sublineage CG23-I. ST86 was the only clone shared by a human and a nonhuman (dog) source. Our work shows the limited transmission of KpSC isolates between human and nonhuman sources and points to the hospital setting as a cornerstone of the spread of MDR clones and antibiotic resistance genes.

IMPORTANCE In this study, we characterized the presence and genomic features of isolates of the *Klebsiella pneumoniae* species complex (KpSC) from human and nonhuman sources in Guadeloupe (French West Indies) in order to identify the reservoirs and routes of transmission. This is the first study in an island environment, an ideal setting that limits the contribution of external imports. Our data showed the limited transmission of KpSC isolates between the different compartments. In contrast, we

Editor Nilton Lincopan, Institute of Biomedical Sciences, Universidade de São Paulo

Copyright © 2022 Dereeper et al. This is an open-access article distributed under the terms of the [Creative Commons Attribution 4.0 International license](https://creativecommons.org/licenses/by/4.0/).

Address correspondence to Sébastien Breurec, sbreurec@gmail.com.

The authors declare no conflict of interest.

Received 17 April 2022

Accepted 8 August 2022

identified the hospital setting as the epicenter of antibiotic resistance due to the nosocomial spread of successful multidrug-resistant (MDR) *K. pneumoniae* clones and antibiotic resistance genes. Ecological barriers and/or limited exposure may restrict spread from the hospital setting to other reservoirs and vice versa. These results highlight the need for control strategies focused on health care centers, using genomic surveillance to limit the spread, particularly of high-risk clones, of this important group of MDR pathogens.

KEYWORDS *Klebsiella pneumoniae*, One Health, ESBL, genomic, Caribbean

Klebsiella pneumoniae is currently recognized as an increasing threat to public health due to the emergence and spread of multidrug-resistant (MDR) isolates associated with hospital outbreaks. Recently, the WHO released a global priority list of antibiotic-resistant bacteria requiring new control strategies, including carbapenem-resistant *K. pneumoniae* (CRKP) and extended-spectrum- β -lactamase (ESBL)-producing *K. pneumoniae* (Kp-ESBL) isolates. ESBL and carbapenemase genes are located on mobile genetic elements and are frequently associated with genes encoding resistance to many other classes of antimicrobial agents, leading to bacteria that are difficult to treat. In Europe, it was estimated that 84,535 cases of infections with CRKP and Kp-ESBL bacteria occurred in 2015 and that these infections accounted for 5,805 attributable deaths (1). Estimations of the burden of MDR *K. pneumoniae* infections are lacking worldwide, but MDR rates are increasing globally. For example, in Senegal, *K. pneumoniae* isolates were the most common bacteria associated with neonatal bloodstream infections, of which 85% were ESBL producers (2). *K. pneumoniae* is also responsible for more severe invasive community-acquired infections, often in healthy individuals, including pyogenic liver abscess, pneumonia, and meningitis (3). Infections are caused by hypervirulent *K. pneumoniae* (HvKp) isolates, which belong to particular clonal groups (CGs). In the past, antimicrobial resistance (AMR) genes and virulence genes were present in specific nonoverlapping genomic lineages, but the frontiers are now blurring, resulting in the emergence of MDR and hypervirulent phenotypes in single *K. pneumoniae* isolates (4).

Phylogenetic studies have revealed that the former *K. pneumoniae* species is a genetically heterogeneous group. It has been redefined as the *Klebsiella pneumoniae* species complex (KpSC), a group of five closely related species distributed into seven phylogroups (Kp1 to Kp7): *K. pneumoniae sensu stricto* (phylogroup Kp1), *K. quasipneumoniae* subsp. *quasipneumoniae* (Kp2), *K. quasipneumoniae* subsp. *similipneumoniae* (Kp4), *K. variicola* subsp. *variicola* (Kp3), *K. variicola* subsp. *tropica* (Kp5), *K. quasivariicola* (Kp6), and *K. africana* (Kp7) (5). In this report, for simplicity, KpSC refers to the *K. pneumoniae* species complex, including the five species/seven phylogroups, and *K. pneumoniae* refers to *K. pneumoniae sensu stricto* (phylogroup Kp1).

In addition to their importance as human pathogens, members of the KpSC can be found in a wide range of ecological niches, such as soil, water, plants, insects, birds, reptiles, and the gut of many mammals (6). However, the prevalence and characteristics of KpSC isolates in these different niches are poorly known due to the lack of large dedicated research efforts. Despite the urgent public health threat now represented by the KpSC, knowledge of the dynamics of the transmission of this bacterial complex from environmental and animal reservoirs to humans using a broad ecological approach with whole-genome sequencing (WGS) is also scarce (7–11).

Guadeloupe, a tropical French overseas territory located in the Caribbean, is considered a very high-resource territory (<https://hdr.undp.org>). Data on the KpSC are scarce and recent for this island. An unusually high prevalence of HvKp (24%) was observed in adults admitted to the intensive care units of two university hospitals in Guadeloupe and Martinique (another French Caribbean territory) for spontaneous community-acquired bacterial meningitis (12). Guadeloupe also faced the emergence of hospital-acquired carbapenemase-producing *K. pneumoniae* infections (13) and an increased incidence of nosocomial Kp-ESBL infections (14, 15). Guadeloupe is an ideal place to study the reservoirs of KpSC isolates and identify the routes of transmission to the

human population due to its insularity, small area (1,436 km²), and small population size (395,700 inhabitants in 2019). The primary objectives of our study were to determine the genomic features of a collection of isolated nonhuman *K. pneumoniae* strains regardless of the putative antibiotic resistance phenotype and to compare them with contemporaneous clinical isolates in order to define the reservoirs of clinical isolates and whether recent transmissions of this pathogen can be detected. The secondary objectives were to determine (i) the prevalence of KpSC members from different sources and (ii) their antibiotic susceptibility patterns.

RESULTS

KpSC isolates, phylogenetic diversity, and AMR. For healthy food-producing animals, a total of 199 fecal samples from 124 pigs and 75 beef cattle were collected from 28 farms and the slaughterhouse (34 additional farms). The prevalences of KpSC isolates were 52.0% in bovines (39/75) and 24.2% in pigs (30/124). KpSC isolates were recovered from all nine poultry farms investigated (11 isolates).

For pets, a single rectal swab was taken from 149 dogs and 73 cats from the main animal shelter of Guadeloupe ($n = 15$) and 7 veterinary clinics ($n = 170$). For the identification of risk factors for the fecal carriage of ESBL-producing KpSC, 37 pets were eliminated because they displayed clinical signs of diarrhea and/or received antibiotic treatment in the previous month. Most of the pets were seen for vaccination ($n = 79$; 42.7%), surgery ($n = 33$; 17.8%), a preventive health visit ($n = 28$; 15.1%), or skin and soft tissue infection ($n = 13$; 7.0%). The rates of fecal carriage of KpSC were 27.4% (20/73) among cats and 49.0% (73/149) among dogs.

Dogs ($P = 0.013$), compared to cats, were significantly associated with KpSC fecal carriage (see Table S1 in the supplemental material).

Considering the environment, totals of 85 locally produced fresh vegetables ($n = 45$), flowering plants ($n = 18$), fruits ($n = 11$ [tomatoes only]), and aromatic herbs ($n = 1$ [thyme]) were collected. The prevalences of KpSC isolates were 90.1% (40/44) in vegetables, 66.6% (12/18) in flowering plants, and 36.4% (4/11) in fruits (tomatoes). The only aromatic herb tested was positive (thyme). Of the 44 raw water samples tested from 29 catchment points, KpSC isolates were detected in 15 samples (34.1%). Totals of 54 soil samples and 21 water samples from rivers or natural ponds located in proximity were investigated at 21 sites located throughout Guadeloupe. The frequencies of detection were 25.9% (14/54) and 33.3% (7/21), respectively. We did not find any significant association between the presence of KpSC isolates and the level of anthropic pressure ($P = 0.795$).

A total of 279 contemporaneous human KpSC isolates were recovered. These bacteria were isolated mainly from urine ($n = 137$; 49.1%), blood ($n = 65$; 23.3%), and wounds ($n = 62$; 22.2%). Six were associated with community liver abscess (10.5% of all community isolates), and 2 were associated with meningitis (3.5%). Of the 222 (79.6%) isolates associated with nosocomial infections, most of them were collected from patients hospitalized in medical wards ($n = 110$; 49.6%), intensive care units ($n = 82$; 36.9%), and emergency units ($n = 62$; 27.9%).

Low levels of resistance to antibiotics were observed for KpSC isolates whatever the source, except for those associated with hospital-acquired infections, which displayed significantly higher rates of resistance rates, whatever the antibiotic tested ($P < 0.027$). Full details of resistance to antibiotics according to the source are available in Table S2. All isolates with decreased phenotypic susceptibility to ertapenem were observed in the hospital setting ($n = 10$), 3 of which displayed carbapenemase genes (2 *bla*_{KPC-2} and 1 *bla*_{NDM-1}). Kp-ESBL isolates (21.4%; 117/546) were associated with hospital-acquired infections only, except for 1 isolate recovered from the feces of one cat. The latter isolate also displayed resistance to fluoroquinolones. Resistance to this major family of antibiotics was not found in any isolate from other animals or the environment. All isolates collected from the environment were assigned a wild-type resistance phenotype, except for one isolate from a vegetable and two from soil (resistance to amoxicillin-

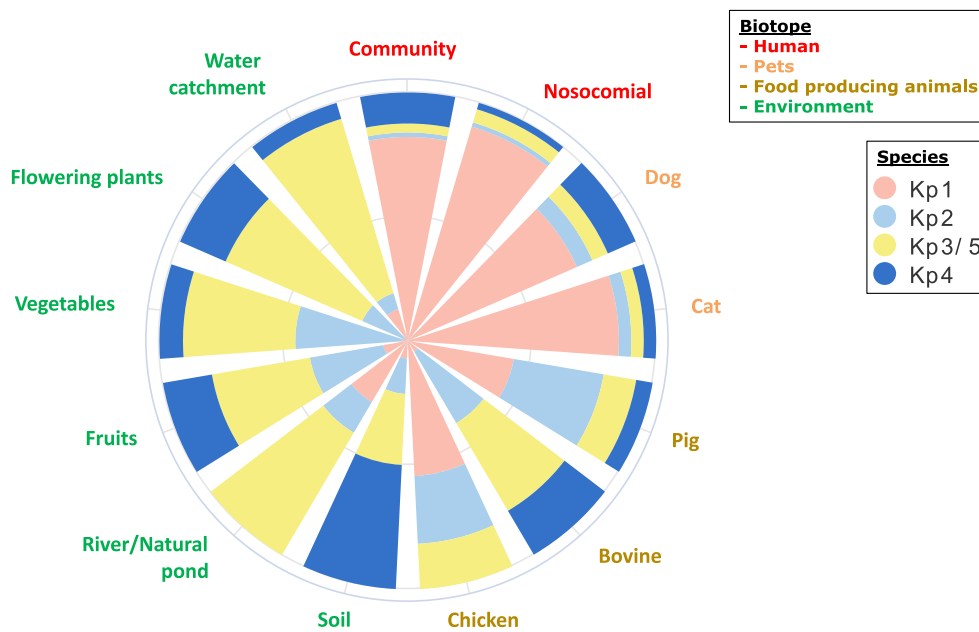


FIG 1 Species and phylogroup distributions of 433 isolates of the *K. pneumoniae* species complex collected in Guadeloupe (French West Indies), according to source. Species were defined by real-time PCR.

clavulanic acid). Globally, the wild-type resistance phenotype corresponded to more than 80% of the isolates detected, except for the hospital setting, where this wild-type phenotype represented 31.8% (Table S2).

Phylogroup typing was performed by real-time PCR (RT-PCR) for all isolates except those recovered from the hospital setting, where ~50% were tested. The 433 KpSC isolates investigated were assigned to *K. pneumoniae* (Kp1 [$n = 245$; 56.6%], Kp2 [$n = 62$; 14.3%], Kp3 or Kp5 [$n = 81$; 18.7%], and Kp4 [$n = 45$; 10.4%]) (Fig. 1 and Table S3). Although the distribution of species and phylogroups across the various sources was clearly nonrandom, each phylogroup was isolated from all sources. A high prevalence of Kp1 and a low prevalence of Kp3 or Kp5 were observed in isolates associated with humans, domestic animals, pigs, and poultry, whereas the opposite relative frequencies of these groups were observed for vegetables, soil/rivers/natural ponds, catchment water, and bovines (Fig. 1 and Table S3). Kp4 was observed mainly in soil ($n = 7$; 50.0%), and Kp2 was observed mainly in food-producing animals (27.3% in poultry, 33.3% in bovines, and 36.7% in pigs) and on vegetables (45.2%). No significant difference was observed when comparing isolates from human and companion animals ($P = 0.092$), those from hospital and community settings ($P = 0.09$), and those from bovines and the environment ($P = 0.79$).

Isolate genomic diversity. WGS was performed on a total of 331 isolates corresponding to a random selection of 66% of those collected from animals and the environment and on 55% of isolates associated with human infections. A full description of the isolates is displayed in Data Set S1 in the supplemental material. Phylogenetic analysis (Fig. 2) was performed to determine the frequencies of Kp1 ($n = 196$; 59.2%), Kp2 ($n = 44$; 13.3%), Kp3 ($n = 50$; 15.1%), and Kp4 ($n = 41$; 12.3%). Assignments were 100% concordant with the phylogrouping results for 433 isolates (see above).

High genetic diversity was found in all sources. A total of 218 sequence types (STs) were identified, with 163 being represented by a single isolate. Nineteen out of 218 STs were high-risk STs (90 isolates; 27.2% of the total), 17 of which were associated with MDR and 2 of which were associated with hypervirulence (ST23 and ST86) (Table 1). All STs represented by 5 or more isolates belonged to previously recognized successful clones. The most frequent STs were high-risk MDR clones, namely, ST17 ($n = 16$; 4.8%), ST307 ($n = 12$; 3.6%), ST11 ($n = 11$; 3.3%), ST147 ($n = 9$; 2.7%), ST152 ($n = 8$; 2.4%), and ST45 ($n = 7$; 2.1%). Most of them were human hospital-acquired Kp-ESBL isolates (47/63; 74.6%).

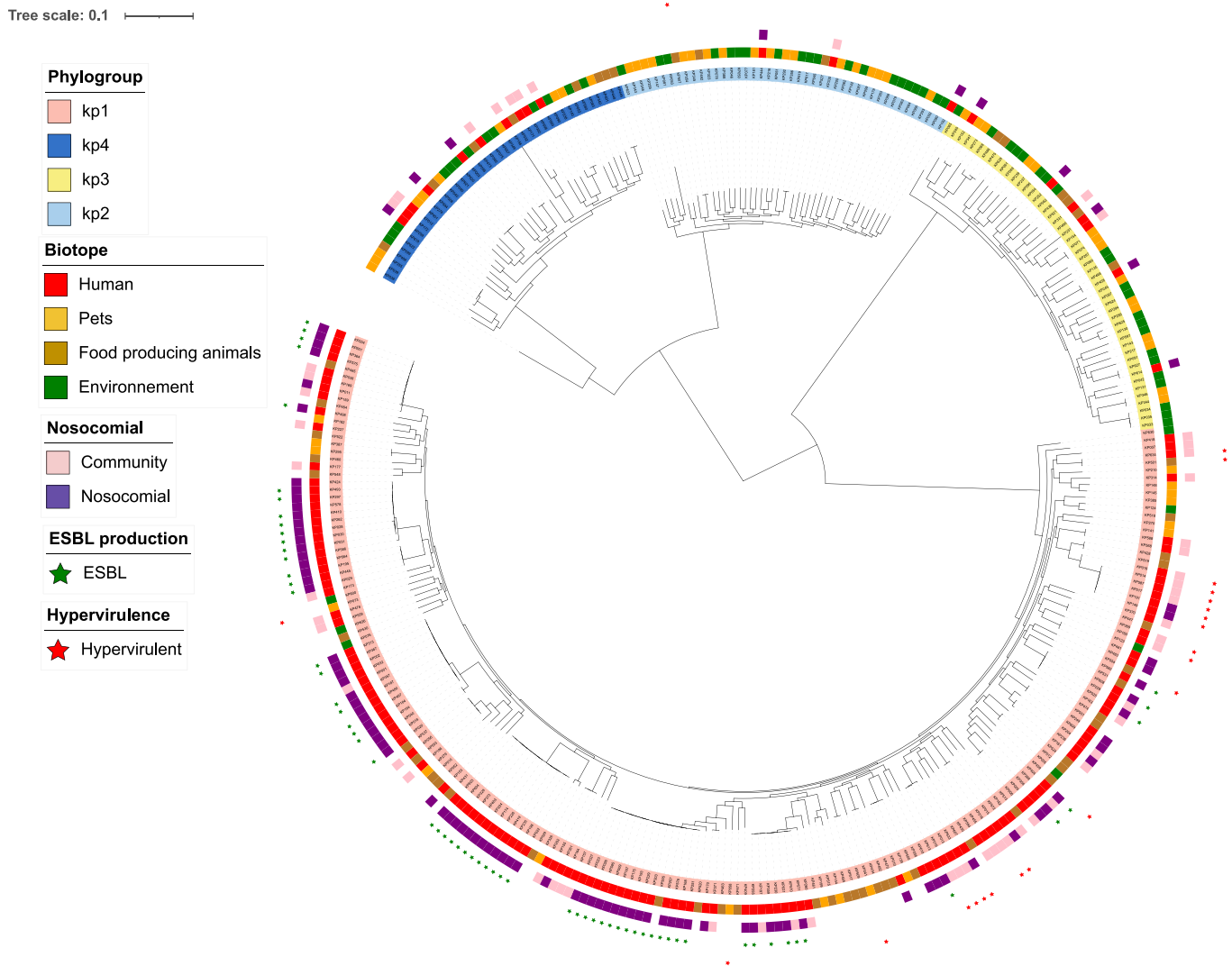


FIG 2 Phylogenetic tree of 331 isolates of the *K. pneumoniae* species complex from Guadeloupe (French West Indies). A maximum likelihood tree was constructed using RAxML based on the core-genome alignment and drawn with iTOL. Leaves are colored according to the phylogroup (Kp1, Kp2, Kp3, and Kp4), and annotation tracks are displayed as follows: source (human/animal/environment), nosocomial/community origin, ESBL production, and hypervirulence.

Only eight STs (five belonged to *K. pneumoniae*, and three belonged to Kp3) were shared between nonhuman and human isolates, with six being strictly shared with pets: three high-risk MDR clones (ST17, ST37, and ST45), one high-risk hypervirulent clone (ST86), and four minor clones (ST2551, ST3600, ST4174, and ST5685). The frequency of STs from clinical isolates also detected in nonhuman samples was 12.7% (8/63). The two isolates assigned to ST45 were common to humans and dogs. They were almost identical (<10 core-genome multilocus sequence typing [cgMLST] allelic mismatches) and were isolated at 1-month intervals (Fig. 3 and Table S4). It should be noted that an ESBL gene (*bla*_{CTX-M-15}) was recovered only in the human isolate. The other shared isolates were genetically more distant. Within the hospital setting, we observed the spread of closely related isolates (<10 single nucleotide polymorphisms [SNPs]) belonging to the four main STs (ST11, ST17, ST45, and ST307) during the 24-month period in different units of the hospital (Fig. 3 and Table S4). For the 4 main STs, isolates from Guadeloupe and those from other Caribbean islands (16) were not directly related based on their genomic sequences (Table S4).

AMR genes and plasmids. In agreement with the phenotypic antibiotic susceptibility patterns, a few isolates (7.9%; 14/178) with a gene(s) or mutation(s) conferring resistance to antimicrobials were recovered outside the human sources (see Table S5 in

TABLE 1 High-risk clones (sequence types) collected from different sources in Guadeloupe (French West Indies)

Clone	No. of clones isolated from source				Total
	Humans	Cats	Dogs	Pigs	
Multidrug resistant					
ST11	11				11
ST13	4				4
ST14	2				2
ST15	1				1
ST17	14	2			16
ST20	2				2
ST29				1	1
ST36	1				1
ST37	1		3		4
ST39	1				1
ST45	6		1		7
ST101	1				1
ST147	9				9
ST152	8				8
ST258	2				2
ST307	12				12
ST392	1				1
Total	76	2	4	1	83
Hypervirulent					
ST23	5				5
ST86	1		1		2
Total	6		1		7

the supplemental material). Out of the 331 isolates investigated, 258 (77.9%) were classified into category 0 (low level of resistance). All of the remaining isolates (resistance scores of 1, 2, and 3 [$n = 73$]) belonged to Kp1 and were isolated mostly from humans ($n = 69$; 94.5%) and the hospital setting ($n = 64$; 87.7%). All isolates carrying an ESBL gene ($n = 61$; 18.4%) were recovered from human isolates associated with nosocomial infections, except for one collected from a cat, in agreement with the results of phenotypic antibiotic susceptibility testing. The ESBL genes were $bla_{CTX-M-15}$ ($n = 59$) and bla_{SHV-12} ($n = 3$). One nosocomial isolate displayed $bla_{CTX-M-15}$ and bla_{SHV-12} . With regard to STs including ≥ 5 isolates, ST17 (15/16), ST147 (7/9), ST152 (5/8), and ST307 (12/12) had high absolute rates of 3rd-generation cephalosporin resistance (3GCR) of over 78%. MDR high-risk clones were associated with a high number of resistance genes, with 11 out of these 13 STs displaying a mean of 6 or more genes.

Two of the three carbapenemase-producing *K. pneumoniae* isolates were sequenced (Kp018 and Kp020); both harbored a bla_{KPC-2} gene, in agreement with PCR typing results (17). They were isolated in the hospital setting. These two isolates were closely related (25 cgMLST allelic mismatches, assigned to the common health care-associated clone ST258). They were classified into the highest-resistance category, harboring the carbapenemase gene bla_{KPC-2} and an alteration of the *mgrB* gene through total (Kp018) or partial (Kp020) deletion, known to confer colistin resistance. They were resistant to colistin (MIC > 4 mg/L). There was no other mechanism of colistin resistance detected in the other isolates, including the acquisition of the *mcr* gene.

Plasmid replicons were found in 67.3% (103/153) of human isolates, 67.8% (38/56) of isolates from pets, 48.3% (29/60) of isolates from food-producing animals, and 45.2% (28/62) of environmental isolates. IncFII (10.6%), IncFIA(HI1) (10.6%), ColRNAI (11.2%), IncR (53.0%), and IncFIB(K) (55.0%) were the most frequently recovered plasmid replicons, with IncFIA (HI1), IncFIB(K), IncFII, and IncR being present in all sources (Table S6).

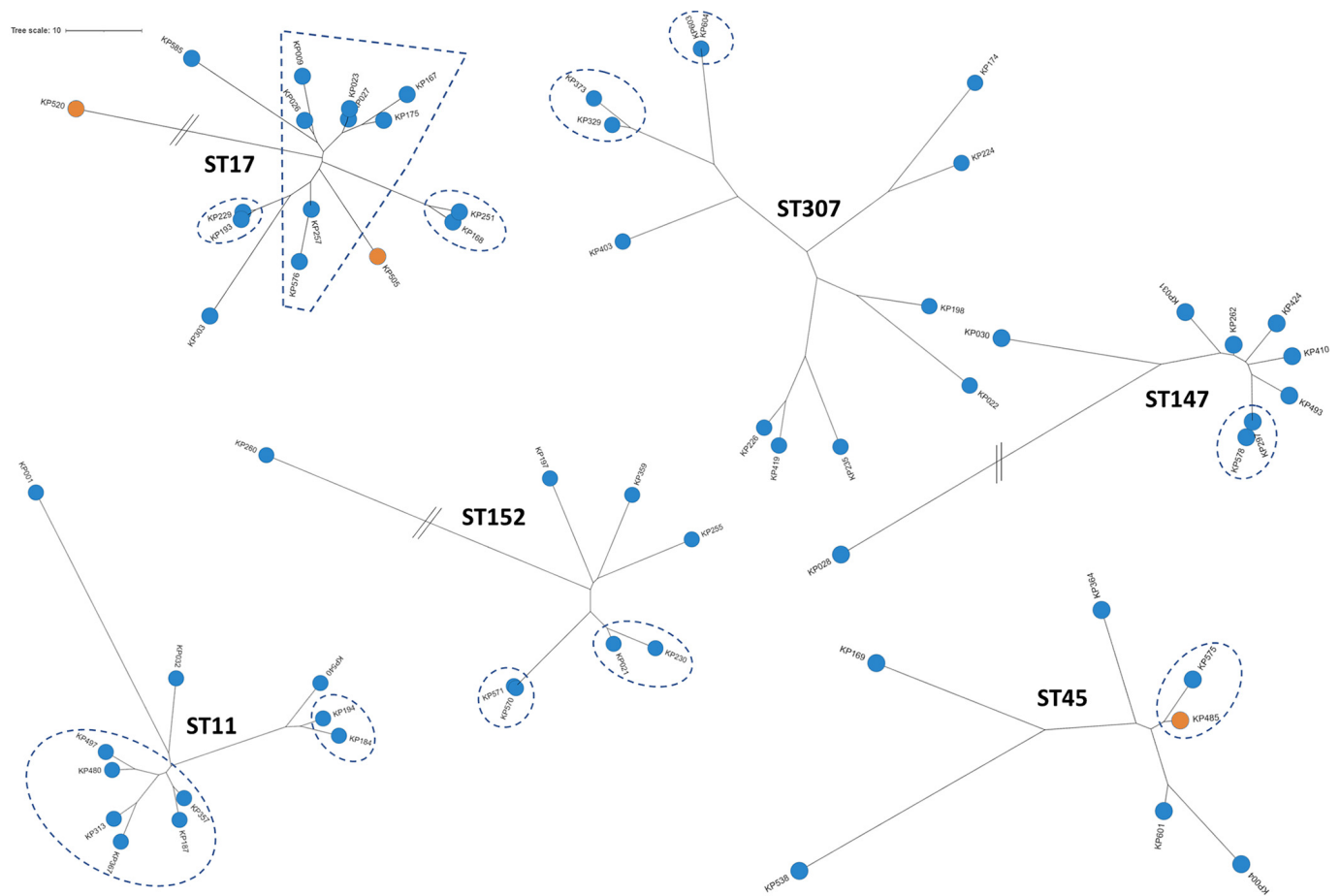


FIG 3 Unrooted phylogenetic trees of the genomes of the six most represented sequence types (STs) recovered in Guadeloupe (French West Indies). All these STs belong to *K. pneumoniae sensu stricto* (Kp1). Isolates are colored according to their source (i.e., human in blue and animal in orange). Using a threshold of <10 SNPs, single strains are framed with dotted lines.

Virulence genes. The prevalences of operons coding for acquired siderophores were 24.2% for yersiniabactin (80/331; 3/80 were incomplete), 6.0% for aerobactin (20/331), and 5.7% for salmochelin (19/331). The genotoxin colibactin loci were present in 4.5% (15/331; 1/15 was incomplete) of the isolates (see Table S7 in the supplemental material). Virulence genes were recovered mainly in human and pet isolates. By comparing human and pet isolates, yersiniabactin (43.8% versus 17.9% [$P = 0.001$]) and colibactin (9.8% versus 0% [$P = 0.013$]) frequencies were significantly higher in humans, but no significant difference was observed for aerobactin (9.2% versus 7.1% [$P = 0.785$]) and salmochelin (5.2% versus 1.9% [$P = 0.450$]). Five isolates from pigs displayed virulence genes, two with yersiniabactin and three with aerobactin, as did one from soil (yersiniabactin). The three aerobactin-positive isolates from pigs carried an *iuc3* aerobactin gene. They came from different farms and belonged to three different genetic backgrounds (ST29, ST432, and ST827). In addition, the *iuc3* gene was recovered on a contig predicted to be plasmidic by combining the MOB-recon and PlasFlow tools. For two out of the three isolates, *iuc3* and an IncFIB(K) replicon were present in the same contig. It should be noted that a presumed plasmid carrying *iuc3* was recovered in one isolate from a dog. The isolate carried an IncFIB(K) replicon but was not recovered on the *iuc3*-positive contig.

In all, a total of 22 isolates (6.6%) were defined as hypervirulent: 15 were associated with community ($n = 12$; 21.4%) and nosocomial ($n = 3$; 3.1%) human infections, 4 (8.9%) were associated with dogs, and 3 (15%) were associated with pigs. The genomic features of HvKp isolates are displayed in Table 2. ST23 (subclade I) was the ST most frequently recovered in human isolates (33.3% of human HvKp isolates). Nine isolates (34.6%) harbored the regulator of mucoid phenotype gene *rmpA*, whereas

TABLE 2 Genomic features of 22 hypervirulent *K. pneumoniae* species complex isolates collected in Guadeloupe (French West Indies) according to source^a

Feature	No. (%) of isolates with feature				
	Humans				
	Community (n = 12)	Nosocomial (n = 3)	Dogs (n = 4)	Pigs (n = 3)	Total (n = 22)
Phylogroup					
<i>K. pneumoniae</i> (Kp1)	12 (100)	3 (100)	3 (75)	3 (100)	21 (95.5)
<i>K. quasipneumoniae</i> subsp. <i>quasipneumoniae</i> (Kp2)	0 (0)	0 (0)	1 (25)	0 (0)	1 (4.5)
Virulence gene					
Colibactin (<i>clb</i>)	10 (83.3)	2 (66.7)	0 (0)	0 (0)	12 (54.5)
Aerobactin (<i>iuc</i>)	12 (100)	3 (100)	2 (50)	3 (100)	20 (90.9)
Salmochelin (<i>iro</i>)	12 (100)	3 (100)	4 (100)	0 (0)	19 (86.4)
<i>rmpA</i>	5 (41.7)	2 (66.7)	2 (50)	0 (0)	9 (40.9)
<i>rmpA2</i>	0 (0)	0 (0)	1 (25)	0 (0)	1 (4.5)
CG and associated ST					
CG5 ST5	0 (0)	0 (0)	1 (25)	0 (0)	1 (4.5)
CG23 ST23	3 (25)	2 (66.7)	0 (0)	0 (0)	5 (22.7)
CG29 ST29	0 (0)	0 (0)	0 (0)	1 (33.3)	1 (4.5)
CG35 ST5750	0 (0)	1 (33.3)	0 (0)	0 (0)	1 (4.5)
CG60 ST60	0 (0)	0 (0)	1 (25)	0 (0)	1 (4.5)
CG65 ST3253	2 (16.7)	0 (0)	0 (0)	0 (0)	2 (9.1)
CG66 ST66	2 (16.7)	0 (0)	0 (0)	0 (0)	2 (9.1)
CG66 ST3252	1 (8.3)	0 (0)	0 (0)	0 (0)	1 (4.5)
CG86 ST86	1 (8.3)	0 (0)	1 (25)	0 (0)	2 (9.1)
CG260 ST260	1 (8.3)	0 (0)	0 (0)	0 (0)	1 (4.5)
CG380 ST380	2 (16.7)	0 (0)	0 (0)	0 (0)	2 (9.1)
CG432 ST432	0 (0)	0 (0)	0 (0)	1 (33.3)	1 (4.5)
CG446 ST446	0 (0)	0 (0)	1 (25)	0 (0)	1 (4.5)
CG827 ST827	0 (0)	0 (0)	0 (0)	1 (33.3)	1 (4.5)
Capsular type					
K1	3 (25)	2 (66.7)	1 (25)	0 (0)	6 (27.3)
K2	7 (58.3)	1 (33.3)	1 (25)	0 (0)	9 (40.9)
Other	2 (16.7)	0 (0)	2 (50)	3 (100)	7 (31.8)

^aCG, clonal group; ST, sequence type.

yersiniabactin, colibactin, and aerobactin operons were present together in 11 (50.0%) isolates. All isolates belonged to *K. pneumoniae*, except for one Kp2 isolate associated with fecal carriage in a dog. These hypervirulence-associated operons were observed mostly in isolates associated with human infections (15/22; 68.2%). The remaining hypervirulent isolates were associated with fecal carriage in dogs ($n = 4$) and pigs ($n = 3$). Human infections corresponded mostly to community-acquired infections (12/15; 80%) and were associated mostly with liver abscess ($n = 5$) and meningitis ($n = 2$). The human hypervirulent isolates were associated mostly with STs typically associated with capsular serotypes K1 ($n = 5$; 22.7%) and K2 ($n = 9$; 40.9%), with the most frequently recovered genetic backgrounds being ST23 ($n = 5$), ST66 ($n = 2$), ST86 ($n = 2$), ST380 ($n = 2$), and ST3253 ($n = 2$). ST23 isolates belonged to the CG23-I lineage (data not shown) according to Lam et al. (18) and carried a yersiniabactin-encoding mobile element (ICEKp10), including genes coding for yersiniabactin and colibactin. All HvKp isolates exhibited the wild-type AMR phenotype.

Virulence genes and MDR convergence. Virulence and antibiotic resistance elements were always observed in distinct isolates, except for one human *K. pneumoniae* isolate (KP586) where a convergence of virulence and MDR was observed. It was assigned to ST392, belonging to the successful MDR clonal group CG147. It was community acquired and displayed the aerobactin virulence operon, *bla*_{CTX-M-15r}, and numerous other resistance genes (*aadA2*, *aac3-IIa*, *bla*_{OXA-1r}, *bla*_{SHV-67r}, *bla*_{TEM-30r}, *catB4*, *dfrA14*, *dfrA32*, *tetA*, *sull*, and *sullI*).

Genome-wide association study. We performed a genome-wide association study (GWAS) on Kp1 isolates in order to identify genes associated with human sources. A genomic region was significantly associated with virulence and host specificity (human). This region includes a cluster of genes located in an integrative conjugative element that mobilizes the *ybt* locus, which encodes the biosynthesis of the siderophore yersiniabactin and its receptor (see Table S8 in the supplemental material). We also found 5 genes significantly associated with human infections encoding proteins with catalytic activity (aconitate hydratase, cytochrome *bo*₃ ubiquinol oxidase, acetoin:2,6-dichlorophenolindophenol oxidoreductase, and dihydrolipoamide dehydrogenase) and 1 encoding a putative methyltransferase. Two genes were significantly associated with nonhuman isolates, including a gene encoding a core component of a type VI secretion system (T6SS) (T6SS baseplate subunit TssK).

DISCUSSION

The increasing level of AMR is a major health hazard for humans and animals (19). Tackling AMR transmission requires investigations of the nonclinical reservoirs and their relative contribution to human infections through the so-called One Health approach. WGS combined with phylogenetic analysis is a powerful approach to provide detailed insights into bacterial transmission dynamics. Despite the urgent public health threat represented by the KpSC, only a few studies so far have identified their genomic features using WGS from nonhuman sources and compared these features with those obtained from contemporaneous and colocalized clinical isolates (7–11). We observed a limited overlap of STs between clinical isolates and local nonhuman isolates, consistent with the results of previous studies (here, 12.7% of human STs; range in previous studies, 5 to 15%). In a large survey of an Italian contemporary KpSC collection from a well-defined geographical region, direct transmission from animal or environmental reservoirs represents a small fraction (<1%) of human infections (20). This highlights the difficulty in identifying direct transmission events for a pathogen characterized by its high genetic diversity without a dominance of specific successful lineages. However, we identify evidence of sporadic transmissions between animals and humans in our set of isolates, such as the presence of 3 hospital-acquired human MDR high-risk *K. pneumoniae* clones (ST17, ST37, and ST45) in pets and 1 successful human hypervirulent *K. pneumoniae* clone (ST86) in a dog. It should be noted that all but one of the STs (one bovine) were shared only between pets and humans, highlighting the potential risk of companion animals in the transmission of KpSC isolates to humans and vice versa, as previously described (21, 22). The presence of the same isolate (<10 SNPs) assigned to ST45 in a dog and in a human collected 1 month apart highlights the importance of the application of basic hygiene rules for contact with companion animals.

Our results illustrate the efficiency of the genomic approach to distinguish epidemiologically related isolates from unrelated ones within hospitals. Genetically closely related isolates were recovered during the 24-month study across different units of the hospital, suggesting an environmental reservoir and long-term transmission. Most nosocomial isolates belonged to high-risk MDR genetic backgrounds, namely, ST11, ST17, ST45, and ST307. These clones have emerged as important vehicles for the worldwide dissemination of AMR determinants (23, 24), including in the Caribbean islands (16).

Although members of the KpSC can be found in a large variety of ecological niches (25), knowledge of the prevalence and distribution of species and phylogroups belonging to this complex according to the source is limited due to the lack of large-scale systematic sampling efforts. Our findings show that the KpSC is ubiquitous, as shown previously in the environment (25), food (26), and the intestines of mammals (21, 27, 28). To the best of our knowledge, dogs were significantly associated for the first time with a higher risk of KpSC carriage than cats. Consistent with the results of previous studies (8, 20), species and phylogroups were not randomly distributed. In particular, *K. pneumoniae sensu stricto* (Kp1) and Kp3 (*K. variicola* subsp. *variicola*) were clearly separated according to ecological niche:

a high prevalence of Kp1 and a low prevalence of Kp3 in humans and domestic and food-producing animals (except for bovines) and the contrary in vegetables and the environment. This may imply ecological barriers that limit the spread of clones and antibiotic resistance genes. However, the distributions of species/phylogroups within the same source were not completely consistent across studies (8, 20), highlighting the need for further work with the inclusion of isolates from wider environmental and animal sources from various geographical areas. For example, we observed a high prevalence of Kp3 among our bovine isolates, probably due to transient flora related to the consumption of raw plants rather than a specific adaptation to colonize their intestine, consistent with the results of a previous study (8) but not with the results of another one (20).

As KpSC isolates were largely recovered in human, animal, and environmental reservoirs in Guadeloupe, we hypothesize that the KpSC could be a major vector for the amplification and spread of antibiotic resistance genes due to its abilities to move between ecological niches, capture and maintain plasmids carrying AMR genes for a long time, and transfer plasmids within KpSC members but also to other important Gram-negative bacteria (6). However, a high level of resistance to antibiotics was rarely found in isolates collected outside the hospital setting but also in isolates of species other than *K. pneumoniae*, illustrated by the almost exclusive presence of ESBL and carbapenemase genes in human *K. pneumoniae* isolates, in agreement with the results of previous studies in Guadeloupe (29–31). Nevertheless, further studies using a shotgun metagenomics approach are needed to access the whole resistome in different ecosystems. Despite the risk of the occasional emergence of novel resistance mechanisms in KpSC isolates from environmental sources, our findings strongly support that the nosocomial setting is central to KpSC resistance dissemination, as observed in Europe for carbapenemase-producing *K. pneumoniae* (32), and that KpSC resistance circulates less frequently between the different compartments.

The *bla*_{CTX-M-15} ESBL gene, first detected in 1999 in India, was the most widely distributed ESBL gene in our set of isolates, in agreement with the worldwide situation. AMR genes and plasmids are often associated with certain *K. pneumoniae* genetic lineages, as highlighted by the success of *K. pneumoniae* ST258 being intricately linked with *bla*_{KPC} (33). Further work will be done to characterize the plasmids carrying ESBL genes using Oxford Nanopore technologies in order to study the dissemination of plasmids and their ESBL genes within the hospital setting.

High levels of virulence also tended to be rare in species other than *K. pneumoniae*. Most of the human HvKp isolates belonged to *K. pneumoniae* ST23 and more precisely to sublineage CG23-I, which emerged in approximately 1928 following the acquisition of ICEKp10 (encoding yersiniabactin and colibactin) and then spread worldwide (18). The uncommon ST66 genetic background (34) was also identified. A reservoir of HvKp was not found outside humans, but dogs could be an important link in the chain of the transmission of this pathogen (9% of hypervirulent isolates were recovered from this species), as highlighted by the presence of ST86 associated with a dog and a human meningitis case. Aerobactin (*iuc* operon) was present in 90% of our HvKp isolates. It is considered a critical siderophore system of HvKp as it contributes predominantly to hypervirulence in laboratory experiments and mouse models of disease, while the inactivation of other siderophore systems has minimal effects (35, 36). Surprisingly, a high frequency of aerobactin (15%) in *K. pneumoniae* isolates from pigs was observed, as previously described in Thailand (37) and Germany (11), probably reflecting an adaptation conferred by this siderophore to porcine hosts. Although our isolates were isolated from pigs from different farms and belonged to different STs, they harbored an *iuc3* gene carried by an IncFIB(K) plasmid (for at least two isolates), as observed in Germany (11). These observations suggest that successful IncFIB(K)/*iuc3*-carrying plasmids have spread across wide geographical distances and occur in different *K. pneumoniae* lineages associated with domestic pigs. The potential risk to animal and human health should also be investigated. Unsurprisingly, a significant association was observed between human infections and the *ybt* locus encoding yersiniabactin,

the most common *K. pneumoniae* high-virulence factor, present in around one-third of clinical isolates (8, 38). The emergence of potentially high-risk MDR and hypervirulent lineages within the hospital setting in Guadeloupe should be monitored, as illustrated by the acquisition of virulence genes in an MDR genetic background (ST392), even if still rarely observed. For the other five genes found to be significantly associated with human infections by GWASs, we did not find any explanation.

Our study represents a large contemporaneous and colocalized sampling and sequencing effort on an island characterized by its small population and area and known to be a hot spot for the spread of health care-associated MDR *K. pneumoniae*. Despite its insular character, which is expected to promote mainly transmissions with local clones and a restricted contribution from the outside, we found limited evidence for direct transmission between human and nonhuman sources (animals and the environment). In contrast, the nosocomial context seems to be a cornerstone of the dissemination of MDR clones and antibiotic resistance genes.

MATERIALS AND METHODS

Fecal samples from healthy food-producing animals and pets. The study design and methods for selecting healthy food-producing animals were described previously (39). Briefly, between January 2018 and May 2019, fecal samples from pigs and beef cattle were collected randomly just after excretion. Fecal material from 17 hen houses (representing 53,000 poultry) was sampled by walking on litter approximately 100 m around a flock in boot socks. In all, the animals originated from 11 pig farms, 8 beef cattle farms, and 9 poultry farms distributed throughout the island and from the only slaughterhouse in Guadeloupe (34 additional farms) for cattle and pigs. Sixty-four percent of farmers declared antibiotic use during the previous 1 year for curative treatment, with the most commonly used antibiotic being tetracycline (69.0%). No ethics committee approval was necessary as no invasive procedure was conducted on live animals according to French national law for the protection of animals (no. 2013-118), which reproduces European directive 2010/63/EU on the protection of animals used for experimental and other scientific purposes.

For pets, from June to September 2019, a single rectal swab was taken from dogs and cats. The animals were included from the main animal shelter of Guadeloupe and seven veterinary clinics located throughout the territory, among animals sent for preventive health services, vaccination, or medical consultation. With regard to the identification of risk factors for the fecal carriage of ESBL-producing KpSC, pets with clinical signs of diarrhea and/or with antibiotic treatment in the previous month were excluded. The information collected for each animal included age, place of residence, general health, and lifestyle (indoors or wandering free outdoors and close contact with other animals or not). The project was approved by the Committee for Ethics in Animal Experiments of the French West Indies and Guyana (reference no. HC_2020_1).

Fresh fruits, fresh vegetables, flowering plants, aromatic herbs, and water and soil samples. Locally produced fresh vegetables, flowering plants, fruits, and aromatic herbs were collected aseptically at four local markets during eight campaigns from January to June 2018. The collected samples spanned producers from 14 municipalities throughout Guadeloupe. Data related to the market and the farm of origin were recorded. When the farm of origin was unknown, multiple vendors in different parts of the market were selected to minimize the likelihood that samples came from the same source.

Raw water samples were collected during the same period at 29 catchment points in 11 municipalities, in partnership with the regional health agency and the hygiene laboratory of the Pasteur Institute of Guadeloupe. These samples corresponded to drinking water before treatment.

From October to December 2019, soil was sampled near rivers and natural ponds: the surface layer (0 to 10 cm) was collected after the removal of plants, pebbles, and conspicuous roots. At each site, two or three samples of soil were taken at least 3 m apart. One sample of water from rivers or natural ponds was collected in proximity. Sampling sites were classified by Q-GIS software into two groups according to their degree of anthropogenic pressure: (i) wilderness with no human presence or countryside with limited human activities, (ii) human-perturbed landscapes with a matrix of agriculture and livestock activities, and (iii) urban and suburban areas with high levels of human activity.

All samples were kept at 2°C to 8°C and processed at the laboratory within 24 h.

Clinical isolates. Between January 2018 and December 2019, 279 contemporaneous KpSC isolates were collected from patients admitted to the University Hospital of Guadeloupe, a 900-bed teaching hospital. Isolation was performed as part of the routine activity of the hospital bacteriological diagnostic laboratory. All presumptive HvKp isolates, defined as KpSC isolates associated with community-acquired monomicrobial liver abscess or other monomicrobial invasions of normally sterile sites (e.g., meningitis), were included in the same period. The following metadata were anonymously recorded: date of hospital admission, ward of hospitalization, date and site of sampling, and antimicrobial susceptibility testing results. Isolates were considered to be community acquired if they were recovered by culture from a sample obtained within 48 h after admission in a patient with no risk factors for nosocomial acquisition in the previous year, namely, hospitalization or surgery, the use of an indwelling catheter or a percutaneous device, or frequent exposure to health care facilities for an underlying chronic disorder. All other

isolates were considered to be hospital acquired. The study protocol was approved by the ethic committee of the University Hospital of Guadeloupe (reference no. A5_19_12_05_TRAMID).

K. pneumoniae species complex isolation and antimicrobial susceptibility testing. Fruits, vegetables, flowering plants, and aromatic herbs after mixing and soil and fecal samples (stool and boot sock samples and endorectal swabs) were inoculated into Luria-Bertani (LB) broth with amoxicillin (10-mg/L final concentration). For water samples, 100 mL of serially diluted samples was filtered through a 0.45- μ m membrane filter (Millipore, Guyancourt, France), and the membranes were placed into 9 mL of LB broth with amoxicillin (10-mg/L final concentration). After incubation for 18 h at 37°C, 100 μ L of the enrichment culture was plated onto two selective media, Simmons citrate agar (Becton, Dickinson, USA) with 1% inositol (SCAI) medium agar plates for KpSC detection and chromogenic agar (CCA) with ceftriaxone at 4 mg/L (CHROMagar) for the detection of 3GCR KpSC. Presumptive Enterobacteriaceae colonies on selective SCAI medium (large, yellow, glossy colonies) and selective CCA with ceftriaxone (pink colonies), corresponding to oxidase-negative and Gram-negative bacilli, were isolated randomly and identified by matrix-assisted laser desorption ionization–time of flight mass spectrometry (MALDI-TOF MS) on an Axima Performance system (Shimadzu Corp., Japan).

Three colonies were identified randomly for each identical morphology. Susceptibility to amoxicillin (10 μ g), amoxicillin-clavulanic acid (20 μ g/10 μ g), ticarcillin (75 μ g), cefotaxime (5 μ g), ceftazidime (10 μ g), cefepime (30 μ g), ceftoxitin (30 μ g), aztreonam (30 μ g), ertapenem (10 μ g), gentamicin (15 μ g), amikacin (30 μ g), trimethoprim-sulfamethoxazole (1.25/23.75 μ g), nalidixic acid (30 μ g), and ciprofloxacin (5 μ g) was tested by the disk diffusion method on Mueller-Hinton agar (Bio-Rad, Marnes-la-Coquette, France), and the production of ESBL was detected by the double-disk synergy test, according to 2020 guidelines of CA-SFM/EUCAST (<https://www.sfm-microbiologie.org/2020/10/02/casfm-eucast-v1-2-octobre-2020/>). Isolates with a resistant or intermediate phenotype were classified together for analysis. Growth inhibition diameters were measured with the Adagio automated system (Bio-Rad, Marnes-la-Coquette, France). Susceptibility to colistin was determined using a MicroScan Walkaway Plus system (Beckman Coulter, USA). Resistant isolates were defined by an MIC of >4 mg/L. If more than one KpSC isolate with the same antibiotic susceptibility pattern was isolated from the same sample, only the first one was analyzed. An MDR KpSC isolate was defined as an isolate resistant to three or more antimicrobial classes (40).

DNA extraction, K. pneumoniae species complex identification, and carbapenemase resistance gene screening. DNA was extracted with a DNA minikit (Qiagen, Germany). A real-time PCR method based on specific sets of primers and probes was used to identify *K. pneumoniae* isolates to the species and phylogroup levels. The protocols are provided in Text S1 in the supplemental material. All isolates were handled using this method, except for those associated with nosocomial infections, of which about half were randomly selected due to the large number of isolates collected. For isolates with decreased susceptibility to ertapenem, carbapenemase genes were searched for using PCR amplification according to methods described in a previous study (17).

Genome sequencing and data analysis. WGS was carried out at the Plateforme de Microbiologie Mutualisée of the Institut Pasteur (Paris, France). Reads were trimmed and filtered with AlienTrimmer software (41), yielding a mean estimated coverage of 86-fold. Genomic assemblies were performed using SPAdes software (42), and the quality of the assembly was evaluated using QUAST software (43). Genomes with a cumulative size of contigs of >6 Mb (expected size of ~ 5.5 Mb) or a number of contigs of >500 were discarded, as we suspected the presence of multiple clones. The mean N_{50} was 197,864 bp.

KpSC phylogrouping, sequence type (ST) assignment, and antibiotic resistance gene detection were performed using Kleborate (44). This tool classified isolates according to the content of resistance genes as follows: 0 for no ESBL and no carbapenemase, 1 for ESBL positive, 2 for carbapenemase positive, and 3 for carbapenemase with colistin resistance (44). Plasmid replicons were identified using the PlasmidFinder database available from ABRicate software, using a minimum coverage and a minimum identity of 90% (45). Kleborate was also used to search for yersiniabactin, colibactin, aerobactin, and salmochelin operons and the presence of *rmpA* and *rmpA2* and to predict capsular types (44). BIGSdb (<https://bigsdbs.pasteur.fr/klebsiella/>) was used to check for the presence of intact *iucABCD-iutA* (aerobactin) and *iroBCDN* (salmochelin) operons. To study the genetic support of *iuc* genes in more detail, the corresponding contigs were predicted to be plasmid or chromosomally associated by combining 2 different software tools (MOB-recon and PlasFlow) with default thresholds (46, 47).

The definitions described previously by Huynh et al. (48) for hypervirulent isolates and successful clones were used. Hypervirulent isolates were defined as isolates harboring at least one of the *rmpA* and *rmpA2* genes and/or at least one complete operon among *iucABCD-iutA* and *iroBCDN*. Successful clones were defined as those belonging to an ST represented at least 10 times in NCBI genomes and mentioned in the title or abstract of at least five publications in NCBI PubMed ("*Klebsiella*" + "*pneumoniae*" + "STxxx").

Pangenome and phylogenetic analyses. Filtered raw contigs were assigned to the chromosome or plasmid using MOB-suite (47), and chromosomal contigs were then ordered and oriented using RaGOO (49) with a publicly available complete assembled genome available for each *Klebsiella* species, according to the species assigned previously by Kleborate (44). Genome annotation was performed using Prokka (50), and pangenome analysis was performed using Roary software (51). Pangenome matrix representation was done using the Roary plots python utility (https://github.com/sanger-pathogens/Roary/tree/master/contrib/roary_plots).

Based on the core-genome alignment provided by Roary and after the identification of recombinant regions and the reduction of the alignment using ClonalFrameML (52), a global phylogenetic tree was constructed using RAxML (53) and visualized using iTOL (54). Estimation of the number of single nucleotide polymorphisms (SNPs) between isolates assigned to an identical ST was performed using the

PathogenWatch platform (<https://pathogen.watch>), an online global database for genomic surveillance of KpSC isolates (55). Contigs were uploaded, and collections were created separately by ST, in order to generate SNP difference matrix files on the one hand and Newick files to be visualized as unrooted trees using iTOL on the other hand. Two isolates were considered clonal when the number of SNPs between them was <10. Additional *Klebsiella* strains obtained from other Caribbean islands (16) were added to the collection. For this purpose, raw FASTQ files were retrieved from the SRA and preliminarily assembled using Unicycler to then be uploaded to PathogenWatch.

The pangenome matrix from Roary consists of gene presence or absence for each genome. It was used as the input for Scoary V1.6.16 (56) in order to search for genes associated with humans or other sources. Due to the bias that might be introduced by the high predominance of Kp1 in human isolates, we focused our pangenome-wide association studies (pan-GWASs) on Kp1 only. Genes returning a Bonferroni-corrected *P* value of ≤ 0.05 were considered to be significantly present/absent and were further investigated.

Statistical analyses. Results were expressed as numbers and frequencies. In bivariate analyses, the χ^2 test (or Fisher's exact test when appropriate) was used to compare categorical data between groups. A logistic regression model was performed to identify factors associated with the presence of KpSC isolates and to calculate crude and adjusted odds ratios and their 95% confidence intervals. Factors with a *P* value of <0.20 in the bivariate analysis were retained for the multivariate analysis. For all tests, we considered a *P* value of <0.05 to be significant. Statistical analyses were performed using SPSS (V21; IBM SPSS Statistics, Chicago, IL).

Data availability. Reads were deposited in the NCBI SRA public archives under BioProject accession no. [PRJNA778230](https://www.ncbi.nlm.nih.gov/bioproject/PRJNA778230).

SUPPLEMENTAL MATERIAL

Supplemental material is available online only.

SUPPLEMENTAL FILE 1, PDF file, 1.4 MB.

SUPPLEMENTAL FILE 2, XLSX file, 0.03 MB.

SUPPLEMENTAL FILE 3, XLS file, 0.03 MB.

SUPPLEMENTAL FILE 4, XLSX file, 0.2 MB.

ACKNOWLEDGMENTS

We thank the Plateforme de Microbiologie Mutualisée (P2M) of the Institut Pasteur for Illumina sequencing. We thank all of the students, technicians, and clinicians involved in this work at the Pasteur Institute of Guadeloupe and the University Hospital Center of Guadeloupe.

Sylvain Brisse and Sébastien Breurec conceived and designed the study. Gaëlle Gruel, Matthieu Pot, Stéphanie Guyomard-Rabenirina, Sylvaine Bastian, Moana Gelu-Simeon, Séverine Ferdinand, Marc Valette, and Sébastien Breurec collected biological samples, isolates, and epidemiological data. Elodie Barbier, Carla Rodrigues, Pascal Piveteau, and Sylvain Brisse provided laboratory support and validated the analysis of the RT-PCR assays. Virginie Passet and Sylvain Brisse performed the MLST and cgMLST curation/analyses. Alexis Dereeper, Pascal Piveteau, Yann Reynaud, Sylvain Brisse, and Sébastien Breurec analyzed the data. Benoit Tressieres performed the statistical analyses. Alexis Dereeper and Sébastien Breurec wrote the initial version of the manuscript. All authors provided input to the manuscript and reviewed the final version.

We declare that there are no conflicts of interest.

This work was supported by a FEDER grant financed by the European Union and Guadeloupe Region (Programme Opérationnel FEDER-Guadeloupe-Conseil Régional 2014–2020, grant no. 2015-FED-192). Carla Rodrigues was supported financially by the MedVetKlebs project, a component of the One Health European Joint Programme (EJP), which has received funding from the European Union's Horizon 2020 Research and Innovation Programme under grant agreement no. 773830, and by a Pasteur-Roux fellowship from the Institut Pasteur.

REFERENCES

1. Cassini A, Högberg LD, Plachouras D, Quattrocchi A, Hoxha A, Simonsen GS, Colomb-Cotinat M, Kretzschmar ME, Devleeschauwer B, Cecchini M, Ouakrim DA, Oliveira TC, Struelens MJ, Suetens C, Monnet DL, Burden of AMR Collaborative Group. 2019. Attributable deaths and disability-adjusted life-years caused by infections with antibiotic-resistant bacteria in the EU and the European Economic Area in 2015: a population-level modelling analysis. *Lancet Infect Dis* 19:56–66. [https://doi.org/10.1016/S1473-3099\(18\)30605-4](https://doi.org/10.1016/S1473-3099(18)30605-4).

2. Breurec S, Bouchiat C, Sire JM, Moquet O, Bercion R, Cisse MF, Glaser P, Ndiaye O, Ka S, Salord H, Seck A, Sy HS, Michel R, Garin B. 2016. High third-generation cephalosporin resistant Enterobacteriaceae prevalence rate among neonatal infections in Dakar, Senegal. *BMC Infect Dis* 16:587. <https://doi.org/10.1186/s12879-016-1935-y>.
3. Shon AS, Bajwa RP, Russo TA. 2013. Hypervirulent (hypermucoviscous) *Klebsiella pneumoniae*: a new and dangerous breed. *Virulence* 4:107–118. <https://doi.org/10.4161/viru.22718>.
4. Wyres KL, Wick RR, Judd LM, Froumine R, Tokolyi A, Gorrie CL, Lam MMC, Duchene S, Jenney A, Holt KE. 2019. Distinct evolutionary dynamics of horizontal gene transfer in drug resistant and virulent clones of *Klebsiella pneumoniae*. *PLoS Genet* 15:e1008114. <https://doi.org/10.1371/journal.pgen.1008114>.
5. Rodrigues C, Passet V, Rakotondraso A, Diallo TA, Criscuolo A, Brisse S. 2019. Description of *Klebsiella africanensis* sp. nov., *Klebsiella variicola* subsp. *tropicalensis* subsp. nov. and *Klebsiella variicola* subsp. *variicola* subsp. nov. *Res Microbiol* 170:165–170. <https://doi.org/10.1016/j.resmic.2019.02.003>.
6. Wyres KL, Holt KE. 2018. *Klebsiella pneumoniae* as a key trafficker of drug resistance genes from environmental to clinically important bacteria. *Curr Opin Microbiol* 45:131–139. <https://doi.org/10.1016/j.mib.2018.04.004>.
7. Davis GS, Waits K, Nordstrom L, Weaver B, Aziz M, Gauld L, Grande H, Bigler R, Horwinski J, Porter S, Stegger M, Johnson JR, Liu CM, Price LB. 2015. Intermingled *Klebsiella pneumoniae* populations between retail meats and human urinary tract infections. *Clin Infect Dis* 61:892–899. <https://doi.org/10.1093/cid/civ428>.
8. Holt KE, Wertheim H, Zadoks RN, Baker S, Whitehouse CA, Dance D, Jenney A, Connor TR, Hsu LY, Severin J, Brisse S, Cao H, Wilksch J, Gorrie C, Schultz MB, Edwards DJ, Nguyen KV, Nguyen TV, Dao TT, Mensink M, Minh VL, Nhu NT, Schultsz C, Kuntaman K, Newton PN, Moore CE, Strugnell RA, Thomson NR. 2015. Genomic analysis of diversity, population structure, virulence, and antimicrobial resistance in *Klebsiella pneumoniae*, an urgent threat to public health. *Proc Natl Acad Sci U S A* 112:E3574–E3581. <https://doi.org/10.1073/pnas.1501049112>.
9. Ludde C, Moradigaravand D, Jamrozny D, Gouliouris T, Blane B, Naydenova P, Hernandez-Garcia J, Wood P, Hadjirin N, Radakovic M, Crawley C, Brown NM, Holmes M, Parkhill J, Peacock SJ. 2020. A One Health study of the genetic relatedness of *Klebsiella pneumoniae* and their mobile elements in the east of England. *Clin Infect Dis* 70:219–226. <https://doi.org/10.1093/cid/ciz174>.
10. Runcharoen C, Moradigaravand D, Blane B, Paksanont S, Thammachote J, Anun S, Parkhill J, Chantratita N, Peacock SJ. 2020. Whole genome sequencing reveals high-resolution epidemiological links between clinical and environmental *Klebsiella pneumoniae*. *Genome Med* 9:6. <https://doi.org/10.1186/s13073-017-0397-1>.
11. Klaper K, Hammerl JA, Rau J, Pfeifer Y, Werner G. 2021. Genome-based analysis of *Klebsiella* spp. isolates from animals and food products in Germany, 2013–2017. *Pathogens* 10:573. <https://doi.org/10.3390/pathogens10050573>.
12. Piednoir P, Clarac U, Rolle A, Bastian S, Gruel G, Martino F, Mehdaoui H, Valette M, Breurec S, Carles M. 2020. Spontaneous community-acquired bacterial meningitis in adults admitted to the intensive care units in the Caribbean French West Indies: unusual prevalence of *Klebsiella pneumoniae*. *Int J Infect Dis* 100:473–475. <https://doi.org/10.1016/j.ijid.2020.09.1420>.
13. Bastian S, Nordmann P, Creton E, Malpote E, Thiery G, Martino F, Breurec S, Dortet L. 2015. First case of NDM-1 producing *Klebsiella pneumoniae* in Caribbean islands. *Int J Infect Dis* 34:53–54. <https://doi.org/10.1016/j.ijid.2015.03.002>.
14. Arnaud I, Maugat S, Jarlier V, Astagneau P, National Early Warning, Investigation and Surveillance of Healthcare-Associated Infections Network (RAISIN)/Multidrug Resistance Study Group. 2015. Ongoing increasing temporal and geographical trends of the incidence of extended-spectrum beta-lactamase-producing Enterobacteriaceae infections in France, 2009 to 2013. *Euro Surveill* 20:30014. <https://doi.org/10.2807/1560-7917.ES.2015.20.36.30014>.
15. Le Terrier C, Vinetti M, Bonjean P, Richard R, Jarrige B, Pons B, Madeux B, Piednoir P, Ardisson F, Elie E, Martino F, Valette M, Ollier E, Breurec S, Carles M, Thiery G. 2021. Impact of a restrictive antibiotic policy on the acquisition of extended-spectrum beta-lactamase-producing Enterobacteriaceae in an endemic region: a before-and-after, propensity-matched cohort study in a Caribbean intensive care unit. *Crit Care* 25:261. <https://doi.org/10.1186/s13054-021-03660-z>.
16. Heinz E, Brindle R, Morgan-McCalla A, Peters K, Thomson NR. 2019. Caribbean multi-centre study of *Klebsiella pneumoniae*: whole-genome sequencing, antimicrobial resistance and virulence factors. *Microb Genom* 5:e000266. <https://doi.org/10.1099/mgen.0.000266>.
17. Poirel L, Walsh TR, Cuvillier V, Nordmann P. 2011. Multiplex PCR for detection of acquired carbapenemase genes. *Diagn Microbiol Infect Dis* 70:119–123. <https://doi.org/10.1016/j.diagmicrobio.2010.12.002>.
18. Lam MMC, Wyres KL, Duchene S, Wick RR, Judd LM, Gan YH, Hoh CH, Archuleta S, Molton JS, Kalimuddin S, Koh TH, Passet V, Brisse S, Holt KE. 2018. Population genomics of hypervirulent *Klebsiella pneumoniae* clonal-group 23 reveals early emergence and rapid global dissemination. *Nat Commun* 9:2703. <https://doi.org/10.1038/s41467-018-05114-7>.
19. Antimicrobial Resistance Collaborators. 2022. Global burden of bacterial antimicrobial resistance in 2019: a systematic analysis. *Lancet* 399:629–655. [https://doi.org/10.1016/S0140-6736\(21\)02724-0](https://doi.org/10.1016/S0140-6736(21)02724-0).
20. Thorpe T, Booton R, Kallonen T, Gibbon MJ, Couto N, Passet V, Lopez Fernandez JS, Rodrigues C, Matthews L, Mitchell S, Reeve E, David S, Merla C, Corbella M, Ferrari C, Comandatore F, Marone P, Brisse S, Sasser S, Corander J, Feil EJ. 2021. One Health or Three? Transmission modelling of *Klebsiella* isolates reveals ecological barriers to transmission between humans, animals and the environment. *bioRxiv*. <https://doi.org/10.1101/2021.08.05.455249>.
21. Marques C, Belas A, Aboim C, Cavaco-Silva P, Trigueiro G, Gama LT, Pomba C. 2019. Evidence of sharing of *Klebsiella pneumoniae* strains between healthy companion animals and cohabiting humans. *J Clin Microbiol* 57:e01537-18. <https://doi.org/10.1128/JCM.01537-18>.
22. Garcia-Fierro R, Drapeau A, Dazas M, Saras E, Rodrigues C, Brisse S, Madec J-Y, Haenni M. 2022. Comparative phylogenomics of ESBL-, AmpC- and carbapenemase-producing *Klebsiella pneumoniae* originating from companion animals and humans. *J Antimicrob Chemother* 77:1263–1271. <https://doi.org/10.1093/jac/dkac041>.
23. Breurec S, Guessennd N, Timinouni M, Le TAH, Cao V, Ngandjio A, Randrianirina F, Thiberge JM, Kinana A, Dufougeray A, Perrier-Gros-Claude JD, Boisier P, Garin B, Brisse S. 2013. *Klebsiella pneumoniae* resistant to third-generation cephalosporins in five African and two Vietnamese major towns: multiclonal population structure with two major international clonal groups, CG15 and CG258. *Clin Microbiol Infect* 19:349–355. <https://doi.org/10.1111/j.1469-0691.2012.03805.x>.
24. Peirano G, Chen L, Kreiswirth BN, Pitout JDD. 2020. Emerging antimicrobial-resistant high-risk *Klebsiella pneumoniae* clones ST307 and ST147. *Antimicrob Agents Chemother* 64:e01148-20. <https://doi.org/10.1128/AAC.01148-20>.
25. Hu Y, Anes J, Devineau S, Fanning S. 2021. *Klebsiella pneumoniae*: prevalence, reservoirs, antimicrobial resistance, pathogenicity, and infection. A hitherto unrecognized zoonotic bacterium. *Foodborne Pathog Dis* 18:63–84. <https://doi.org/10.1089/fpd.2020.2847>.
26. Rodrigues C, Hauser K, Cahill N, Ligowska-Marzeta M, Centorotola G, Cornacchia A, Garcia Fierro R, Haenni M, Nielsen EM, Piveteau P, Barbier E, Morris D, Pomilio F, Brisse S. 2022. High prevalence of *Klebsiella pneumoniae* in European food products: a multicentric study comparing culture and molecular detection methods. *Microbiol Spectr* 10:e02376-21. <https://doi.org/10.1128/spectrum.02376-21>.
27. Zadoks RN, Griffiths HM, Munoz MA, Ahlstrom C, Bennett GJ, Thomas E, Schukken YH. 2011. Sources of *Klebsiella* and *Raoultella* species on dairy farms: be careful where you walk. *J Dairy Sci* 94:1045–1051. <https://doi.org/10.3168/jds.2010-3603>.
28. Zhong XS, Li YZ, Ge J, Xiao G, Mo Y, Wen YQ, Liu JP, Xiong YQ, Qiu M, Huo ST, Cheng MJ, Chen Q. 2020. Comparisons of microbiological characteristics and antibiotic resistance of *Klebsiella pneumoniae* isolates from urban rodents, shrews, and healthy people. *BMC Microbiol* 20:12. <https://doi.org/10.1186/s12866-020-1702-5>.
29. Guyomard-Rabenirina S, Reynaud Y, Pot M, Albina E, Couvin D, Ducat C, Gruel G, Ferdinand S, Legreneur P, Le Hello S, Malpote E, Sadiikalay S, Talarmin A, Breurec S. 2020. Antimicrobial resistance in wildlife in Guadeloupe (French West Indies): distribution of a single *bla*_{CTX-M-1}/Incl1/ST3 plasmid among humans and wild animals. *Front Microbiol* 11:1524. <https://doi.org/10.3389/fmicb.2020.01524>.
30. Guyomard-Rabenirina S, Weill FX, Le Hello S, Bastian S, Berger F, Ferdinand S, Legreneur P, Loraux C, Malpote E, Muanza B, Richard V, Talarmin A, Breurec S. 2019. Reptiles in Guadeloupe (French West Indies) are a reservoir of major human *Salmonella enterica* serovars. *PLoS One* 14:e0220145. <https://doi.org/10.1371/journal.pone.0220145>.
31. Pot M, Reynaud Y, Couvin D, Ducat C, Ferdinand S, Gravey F, Gruel G, Guerin F, Malpote E, Breurec S, Talarmin A, Guyomard-Rabenirina S. 2021. Wide distribution and specific resistance pattern to third-generation cephalosporins of *Enterobacter cloacae* complex members in humans and in the environment in Guadeloupe (French West Indies). *Front Microbiol* 12:628058. <https://doi.org/10.3389/fmicb.2021.628058>.
32. David S, Reuter S, Harris SR, Glasner C, Feltwell T, Argimon S, Abudahab K, Goater R, Giani T, Errico G, Aspbury M, Sjunnebo S, EuSCAPE Working

- Group, ESGEM Study Group, Feil EJ, Rossolini GM, Aanensen DM, Grundmann H. 2019. Epidemic of carbapenem-resistant *Klebsiella pneumoniae* in Europe is driven by nosocomial spread. *Nat Microbiol* 4: 1919–1929. <https://doi.org/10.1038/s41564-019-0492-8>.
33. Buckner MMC, Saw HTH, Osagie RN, McNally A, Ricci V, Wand ME, Woodford N, Ivens A, Webber MA, Piddock LJV. 2018. Clinically relevant plasmid-host interactions indicate that transcriptional and not genomic modifications ameliorate fitness costs of *Klebsiella pneumoniae* carbapenemase-carrying plasmids. *mBio* 9:e02303-17. <https://doi.org/10.1128/mBio.02303-17>.
 34. Rodrigues C, d'Humieres C, Papin G, Passet V, Ruppe E, Brisse S. 2020. Community-acquired infection caused by the uncommon hypervirulent *Klebsiella pneumoniae* ST66-K2 lineage. *Microb Genom* 6:mgen000419. <https://doi.org/10.1099/mgen.0.000419>.
 35. Russo TA, Olson R, Macdonald U, Maltzer D, Maltese LM, Drake EJ, Gulick AM. 2014. Aerobactin mediates virulence and accounts for increased siderophore production under iron-limiting conditions by hypervirulent (hypermucoviscous) *Klebsiella pneumoniae*. *Infect Immun* 82:2356–2367. <https://doi.org/10.1128/IAI.01667-13>.
 36. Russo TA, Olson R, MacDonald U, Beanan J, Davidson BA. 2015. Aerobactin, but not yersiniabactin, salmochelin, or enterobactin, enables the growth/survival of hypervirulent (hypermucoviscous) *Klebsiella pneumoniae* *ex vivo* and *in vivo*. *Infect Immun* 83:3325–3333. <https://doi.org/10.1128/IAI.00430-15>.
 37. Leangapichart T, Lunha K, Jiwakanon J, Angkittrakul S, Jarhult JD, Magnusson U, Sunde M. 2021. Characterization of *Klebsiella pneumoniae* complex isolates from pigs and humans in farms in Thailand: population genomic structure, antibiotic resistance and virulence genes. *J Antimicrob Chemother* 76:2012–2016. <https://doi.org/10.1093/jac/dkab118>.
 38. Lam MMC, Wick RR, Wyres KL, Gorrie CL, Judd LM, Jenney AWJ, Brisse S, Holt KE. 2018. Genetic diversity, mobilisation and spread of the yersiniabactin-encoding mobile element ICEKp in *Klebsiella pneumoniae* populations. *Microb Genom* 4:e000196. <https://doi.org/10.1099/mgen.0.000196>.
 39. Gruel G, Sellin A, Riveiro H, Pot M, Breurec S, Guyomard-Rabenirina S, Talarmin A, Ferdinand S. 2021. Antimicrobial use and resistance in *Escherichia coli* from healthy food-producing animals in Guadeloupe. *BMC Vet Res* 17:116. <https://doi.org/10.1186/s12917-021-02810-3>.
 40. Magiorakos AP, Srinivasan A, Carey RB, Carmeli Y, Falagas ME, Giske CG, Harbarth S, Hindler JF, Kahlmeter G, Olsson-Liljequist B, Paterson DL, Rice LB, Stelling J, Struelens MJ, Vatopoulos A, Weber JT, Monnet DL. 2012. Multidrug-resistant, extensively drug-resistant and pandrug-resistant bacteria: an international expert proposal for interim standard definitions for acquired resistance. *Clin Microbiol Infect* 18:268–281. <https://doi.org/10.1111/j.1469-0691.2011.03570.x>.
 41. Criscuolo A, Brisse S. 2014. AlienTrimmer removes adapter oligonucleotides with high sensitivity in short-insert paired-end reads. *Commentary on Turner (2014) Assessment of insert sizes and adapter content in FASTQ data from NexteraXT libraries*. *Front Genet* 5:130. <https://doi.org/10.3389/fgene.2014.00130>.
 42. Bankevich A, Nurk S, Antipov D, Gurevich AA, Dvorkin M, Kulikov AS, Lesin VM, Nikolenko SI, Pham S, Pribelski AD, Pyshtkin AV, Sirotkin AV, Vyahhi N, Tesler G, Alekseyev MA, Pevzner PA. 2012. SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. *J Comput Biol* 19:455–477. <https://doi.org/10.1089/cmb.2012.0021>.
 43. Gurevich A, Saveliev V, Vyahhi N, Tesler G. 2013. QUAST: quality assessment tool for genome assemblies. *Bioinformatics* 29:1072–1075. <https://doi.org/10.1093/bioinformatics/btt086>.
 44. Lam MMC, Wick RR, Watts SC, Cerdeira LT, Wyres KL, Holt KE. 2021. A genomic surveillance framework and genotyping tool for *Klebsiella pneumoniae* and its related species complex. *Nat Commun* 12:4188. <https://doi.org/10.1038/s41467-021-24448-3>.
 45. Carattoli A, Zankari E, Garcia-Fernandez A, Voldby Larsen M, Lund O, Villa L, Moller Aarestrup F, Hasman H. 2014. In silico detection and typing of plasmids using PlasmidFinder and plasmid multilocus sequence typing. *Antimicrob Agents Chemother* 58:3895–3903. <https://doi.org/10.1128/AAC.02412-14>.
 46. Krawczyk PS, Lipinski L, Dziembowski A. 2018. PlasFlow: predicting plasmid sequences in metagenomic data using genome signatures. *Nucleic Acids Res* 46:e35. <https://doi.org/10.1093/nar/gkx1321>.
 47. Robertson J, Nash JHE. 2018. MOB-suite: software tools for clustering, reconstruction and typing of plasmids from draft assemblies. *Microb Genom* 4:e000206. <https://doi.org/10.1099/mgen.0.000206>.
 48. Huynh BT, Passet V, Rakotondrasoa A, Diallo T, Kerleguer A, Hennart M, Lauzanne A, Herindrainy P, Seck A, Bercion R, Borand L, Pardos de la Gandara M, Delarocque-Astagneau E, Guillemot D, Vray M, Garin B, Collard JM, Rodrigues C, Brisse S. 2020. *Klebsiella pneumoniae* carriage in low-income countries: antimicrobial resistance, genomic diversity and risk factors. *Gut Microbes* 11:1287–1299. <https://doi.org/10.1080/19490976.2020.1748257>.
 49. Alonge M, Soyk S, Ramakrishnan S, Wang X, Goodwin S, Sedlazeck FJ, Lippman ZB, Schatz MC. 2019. RaGOO: fast and accurate reference-guided scaffolding of draft genomes. *Genome Biol* 20:224. <https://doi.org/10.1186/s13059-019-1829-6>.
 50. Seemann T. 2014. Prokka: rapid prokaryotic genome annotation. *Bioinformatics* 30:2068–2069. <https://doi.org/10.1093/bioinformatics/btu153>.
 51. Page AJ, Cummins CA, Hunt M, Wong VK, Reuter S, Holden MT, Fookes M, Falush D, Keane JA, Parkhill J. 2015. Roary: rapid large-scale prokaryote pan genome analysis. *Bioinformatics* 31:3691–3693. <https://doi.org/10.1093/bioinformatics/btv421>.
 52. Didelot X, Wilson DJ. 2015. ClonalFrameML: efficient inference of recombination in whole bacterial genomes. *PLoS Comput Biol* 11:e1004041. <https://doi.org/10.1371/journal.pcbi.1004041>.
 53. Stamatakis A. 2014. RAXML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics* 30:1312–1313. <https://doi.org/10.1093/bioinformatics/btu033>.
 54. Letunic I, Bork P. 2021. Interactive Tree of Life (iTOL) v5: an online tool for phylogenetic tree display and annotation. *Nucleic Acids Res* 49:W293–W296. <https://doi.org/10.1093/nar/gkab301>.
 55. Argimon S, David S, Underwood A, Abrudan M, Wheeler NE, Kekre M, Abudahab K, Yeats CA, Goater R, Taylor B, Harste H, Muddyman D, Feil EJ, Brisse S, Holt K, Donado-Godoy P, Ravikumar KL, Okeke IN, Carlos C, Aanensen DM, NIHR Global Health Research Unit on Genomic Surveillance of Antimicrobial Resistance. 2021. Rapid genomic characterization and global surveillance of *Klebsiella* using Pathogenwatch. *Clin Infect Dis* 73:S325–S335. <https://doi.org/10.1093/cid/ciab784>.
 56. Brynildsrud O, Bohlin J, Scheffer L, Eldholm V. 2016. Rapid scoring of genes in microbial pan-genome-wide association studies with Scoary. *Genome Biol* 17:238. <https://doi.org/10.1186/s13059-016-1108-8>.

Metagenome reveals caprine abomasal microbiota diversity at early and late stages of *Haemonchus contortus* infection.

Laura A. Montout^{*1} and Jean-Christophe Bambou^{†1}

¹Agroécologie, génétique et systèmes d'élevage tropicaux – Institut National de Recherche pour l'Agriculture, l'Alimentation et l'Environnement – France

Résumé

Haemonchus contortus is one of the most detrimental gastrointestinal nematode parasites for small ruminants, especially in tropics and subtropics. Gastrointestinal nematode and microbiota share the same microhabitat; thus they interact with each other and their host. Metagenomics tools provide a promising way to examine the alterations in the gastric microbial composition induces by gastrointestinal parasites. In this study, we used metagenomics tools to characterize the impact of *H. contortus* infection on the caprine abomasal microbiota at early and late stage of infection and compared it with non-infected control. Our results showed that *H. contortus* infection caused a significant increase in abomasal pH at early (7 days post-infection) and late stage of infection (56 days post-infection). The analysis of alpha and beta diversity showed that the microbiota diversity both in number and in proportion was significantly affected at early and late stage of infection. All microbiota classes are impacted by *H. contortus* infection but Clostridia and Bacteroidia are more concerned. In infected animals, the genera *Prevotella* decreased at 7 and 56 days post-infection. Here we showed that the abomasal microbiota was significantly affected early after *H. contortus* infection, and these changes persist at late stage of the infection.

Mots-Clés: Parasite host response, Metagenomics

*Intervenant

†Auteur correspondant: jean-christophe.bambou@inrae.fr

KaruBioNet: a network and discussion group for a better collaboration and structuring of bioinformatics in Guadeloupe (French West Indies)

David Couvin^{1*}, Alexis Dereeper^{1*}, Damien F. Meyer^{2,3}, Christophe Noroy⁴, Stanie Gaete⁵, Bernard Bhakkan⁶, Nausicaa Pouillet⁷, Sarra Gaspard⁸, Etienne Bezault⁹, Isabel Marcelino¹, Ludovic Pruneau¹⁰, Wilfried Segretier¹¹, Erick Stattner¹¹, Damien Cazenave¹, Maëlle Garnier¹, Matthieu Pot¹, Sébastien Breurec^{1,12,13}, Séverine Ferdinand¹, Benoît Tressières¹³, Jacqueline Deloumeaux^{5,6}, Silvina Gonzalez-Rizzo¹⁰, Yann Reynaud¹, for the KaruBioNet team^{**}

¹Unité Transmission, Réservoir et Diversité des Pathogènes, Institut Pasteur de Guadeloupe, Les Abymes, Guadeloupe, France. ²CIRAD, UMR ASTRE, Petit-Bourg, Guadeloupe, France. ³ASTRE, Univ Montpellier, CIRAD, INRAE, Montpellier, France. ⁴Développement, Analyse, Transfert et Application (DATA), Lamentin, Guadeloupe, France. ⁵Karubiotec Centre de Ressources Biologiques-UF 0216, CHU de la Guadeloupe, Pointe-à-Pitre, France. ⁶Registre des cancers de Guadeloupe, CHU de la Guadeloupe, Pointe-à-Pitre, France. ⁷URZ Recherches Zootechniques, INRAE, 97170 Petit-Bourg (Guadeloupe), France. ⁸Laboratoire COVACHIMM2E EA3592, Université des Antilles, Pointe-à-Pitre, Guadeloupe, France. ⁹UMR BOREA (MNHN, CNRS-7208, IRD-207, Sorbonne Université, UCN, UA), Université des Antilles, Guadeloupe, France. ¹⁰Équipe « Biologie de la mangrove » UMR7205 « ISYEB » MNHN-CNRS-Sorbonne Université-EPHE-UA, UFR SEN Département de Biologie, Université des Antilles, Pointe-à-Pitre, Guadeloupe. ¹¹Laboratoire de Mathématiques Informatique et Applications (LAMIA), Université des Antilles, Guadeloupe, France. ¹²Faculté de Médecine Hyacinthe Bastaraud, Université des Antilles, Pointe-à-Pitre, France. ¹³Centre d'Investigation Clinique Antilles Guyane, Inserm CIC 1424.

*These authors contributed equally to this work.

**A list of other members of the KaruBioNet team is provided in the Acknowledgements.

Corresponding Author: dcouvin@pasteur-guadeloupe.fr

Paper Reference or Website: Couvin et al., 2022, *KaruBioNet: a network and discussion group for a better collaboration and structuring of bioinformatics in Guadeloupe (French West Indies)*, *Bioinform Adv*, doi:

<https://doi.org/10.1093/bioadv/vbac010>

Abstract: *Sequencing and other biological data are now more frequently available and at a lower price. Mutual tools and strategies are needed to analyze the huge amount of heterogeneous data generated by several research teams and devices. Bioinformatics represents a growing field in the scientific community globally. This multidisciplinary field provides a great amount of tools and methods that can be used to conduct scientific studies in a more strategic way. Coordinated actions and collaborations are needed to find more innovative and accurate methods for a better understanding of real-life data. A wide variety of organizations are contributing to KaruBioNet in Guadeloupe (French West Indies), a Caribbean archipelago. The purpose of this group is to foster collaboration and mutual aid among people from different disciplines using a 'one health' approach, for a better comprehension and surveillance of humans, plants or animals' health and diseases. The KaruBioNet network particularly aims to help researchers in their studies related to 'omics' data, but also more general aspects concerning biological data analysis. This transdisciplinary network is a platform for discussion, sharing, training and support between scientists interested in bioinformatics and related fields. Starting from a little archipelago in the Caribbean, we envision to facilitate exchange between other Caribbean partners in the future, knowing that the Caribbean is a region with non-negligible biodiversity which should be preserved and protected. Joining forces with other Caribbean countries or territories would strengthen scientific collaborative impact in the region. Information related to this network can be found at: <http://www.pasteur-guadeloupe.fr/karubionet.html>. Furthermore, a dedicated 'Galaxy KaruBioNet' platform is available at: http://calamar.univ-ag.fr/c3i/galaxy_karubionet.html.*

Keywords bioinformatics, genomics, galaxy, high-performance-computing, training.

The main purpose of this group is to bring together people who share common problems to better collaborate together. The KaruBioNet network aims to improve the development of bioinformatics at the local or regional level for a better understanding and analysis of real-life data. The major common themes shared by laboratories belonging to the network are microbial evolutionary history, antimicrobial resistance, virulence mechanisms, systems biology, genotyping and data science (**Fig. 1**). The KaruBioNet network particularly aims to help researchers in their studies related to metagenomics, proteomics, genomics (or other 'omics'), as well as more general aspects relating to data analysis, integration, and interpretation. In order to structure discussions and

exchanges between researchers, we have implemented different topics (which could evolve in the future): (i) omics sciences; (ii) artificial intelligence and machine learning; (iii) biochemical analyses; (iv) geographic information systems; (v) databases and software development; (vi) biostatistics; and (vii) epidemiology. A video channel was created to disseminate training materials in French as well as presentations or demonstrations related to bioinformatics. Furthermore, a dedicated Galaxy platform [1] allowing to facilitate bioinformatic analyses is available at: http://calamar.univ-ag.fr/c3i/galaxy_karubionet.html. We also benefit from the ‘Exocet’ High-Performance Computing (HPC) facility of the UA to perform computing calculations.

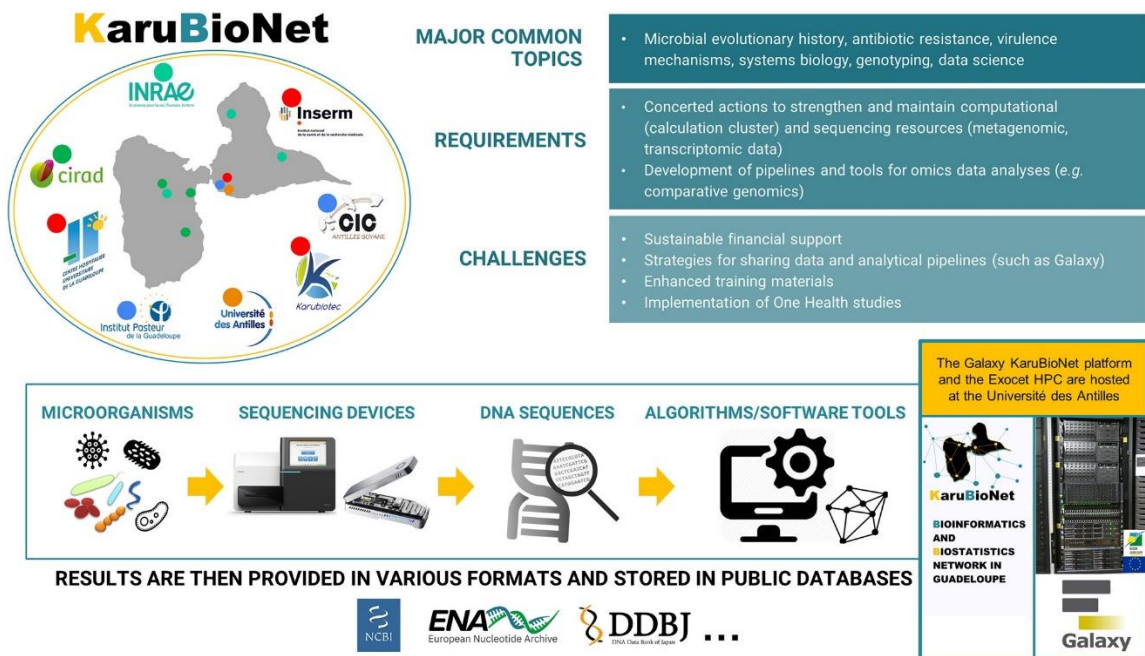


Fig. 1: Mapping showing major common themes shared by research institutions involved in the KaruBioNet. The colored dots represent the location of each institution on the map. This figure also shows a simplified workflow for sequencing data analysis. Resources and platforms are mainly located at the Université des Antilles.

Acknowledgements

A The KaruBioNet network thanks the Regional Council of Guadeloupe for its support. We are also thankful to the ‘Projet MALIN’ consortium (<https://www.projet-malin.fr/>). This work was supported by a FEDER grant financed by the EU. Several computational tests have been performed using Wahoo and Exocet, the clusters of the Centre Commun de Calcul Intensif (C3I) of the Université des Antilles. We are grateful to Nalin Rastogi for helpful discussions. We also thank other people interested or involved in the KaruBioNet team: Raphaël Pasquier, Syndia Sadikalay, Pauline Dentika, Anubis Vega-Rúa, Lyza Hery, Margaux Mulatier, Elodie Calvez, Géliza Gamiette, Antoine Talarmin, Stéphanie Guyomard, Degrâce Batantou, Vincent Guerlais, Mailie Saint-Hilaire, Cécile Martias, Thierry Zozio, Isaure Quétel, Gaëlle Gruel, Nina Allouch, Youri Vingataramin, Marc Romana, Michel Naves, Kizzy-Clara Cita, Daniella Goindin, Alice Choury, Olivier Gros, Larissa Valmy, Jean-Christophe Bambou, Antoine Boullis, Manuel Clergue, Sébastien Regis, Denis Boucaud-Maitre, Christophe Armand, Steve Cériac, Yann Legros, Hugo Boijout, Jean-David Pommier, Jimmy Nagau, Andrei Doncescu, Alain Piétrus, Stéphane Cholet, Jean-Luc Gouridine, Tenissia Cesar, Christopher Cambrone, Georges Minatchy, Suzanne Conjard, Carole Louis-Rose, Elvire Couchy, Murielle Mantran, Vincent Moco, Davy Régalade, Elkana Lesmond, Suly Rambinaising, Mame-Boucar Diouf, Sylvaine Bastian, Raymond Césaire, Benoit Garin, Cécile Herrmann, Chantal Eucar, Charlotte Romero, Kévin Julianus, Kévin Durimel, Lucie Leon, Mounir Serag, Muriel Nicolas, Anthommy Gilles, Dominique Albina, Alex Botino, Betty Fausta, Olivier Watté, Jean-François Dorville, Audrey Robinel, Kenny Chammougou and Éric Tessanne. This initiative was presented at the Caribbean Science and Innovation Meeting in 2019. We also thank Fabien Mareuil, Rémi Planel and Brice Raffestin (Institut Pasteur, Paris), for their help regarding the development of Galaxy. Finally, we are grateful to all present and future partners interested in this project.

References

1. Afgan E. et al. (2018) The Galaxy platform for accessible, reproducible and collaborative biomedical analyses: 2018 update. *Nucleic Acids Res.*, 46, W537–W544. doi: <https://doi.org/10.1093/nar/gky379>

SeBiMER: the bioinformatics platform of Ifremer

Laura Leroi, Cyril Noël, Alexandre Cormier, Pauline Auffret, Alizée Bardon & Patrick Durand

IFREMER-IRSI-SeBiMER, ZI de la Pointe du Diable, 29280 Plouzané, France

Corresponding Author: pgdurand@ifremer.fr

Abstract. *SeBiMER is the bioinformatics platform of Ifremer, the French National Institute for Ocean Science. A team of engineers is in charge of providing a large community of marine biologists with all the requirements (knowledge, software and data) to handle small- and large-scale bioinformatics projects, mostly in three main fields: metabarcoding and metagenomics (eDNA), (meta-)transcriptomics and genomes assembly & annotation. The platform's missions are organized in four poles: (1) IT management: bioinformatics data management and software installation & configuration on DATARMOR supercomputer; (2) bioanalysis: involvement in research projects to provide data analysis expertise; (3) software development: design of new tools and data analysis workflows; (4) training. More: <https://bioinfo.ifremer.fr>.*

Keywords *bioinformatics; environmental DNA; genome assembly; gene expression; workflow.*

The *Service de Bioinformatique de l'Ifremer* (SeBiMER) was created in February 2019 as the bioinformatics platform of Ifremer. It aims at providing a large community of marine biologists with technical and scientific expertise in the three main research domains involving bioinformatics techniques at Ifremer: environmental DNA, gene expression and genomes assembly & annotation. SeBiMER's team is composed of 5 permanent engineers, three of them holding a Ph.D., regularly accompanied by staff on fixed-term contracts. They are involved in many research projects coming from Ifremer and non-Ifremer laboratories, including national (e.g. ATLASea [1]) and international projects (e.g. FAIR-EASE [2]).

The SeBiMER bioinformatics platform provides the following services:

I - IT management. SeBiMER is in charge of centralizing bioinformatics software installation and configuration on DATARMOR supercomputer hosted at Ifremer [3]. All these tools are available from command-line and Ifremer Galaxy web portal. SeBiMER team provides ready-to-use scripts enabling non-expert users to run all these tools through the PBS Pro cluster scheduler. In addition to tool installation, SeBiMER maintains up to date a large collection of public databanks for various analyzes. To achieve these management tasks, the team has developed free and open source softwares, e.g. ToolDirectory to automate the creation of a web-based software catalog [4], BeeDeeM to automate bank installation [5] and OMICS-Catalog workflow to setup a web-based catalog of model genomes under study at Ifremer [6].

II - FAIR data management. SeBiMER and SISMER [7] provide a FAIR procedure to handle bioinformatics data at Ifremer. This procedure is a key element of our Data Management Plan (DMP) and provides a standard life cycle for all sequencing data (short- and long-reads): a formal description of the data project with respect to CoreTrustSeal certification of Ifremer [8], a secure storage on DATARMOR, a formal description of metadata following the ENA metadata model, and, on request, submission to EBI-ENA international databank, and data access through DOI pointing to Ifremer data server [9]. The platform is currently developing athENA, a tool to efficiently collect sequencing project metadata and perform automatic submission to ENA. In this context, SeBiMER takes part in the data brokering work group created as part of MUDIS4LS project directed by the French National Institute for Bioinformatics (IFB) [10].

III - Bioanalysis expertise. Being part of a marine research institute, SeBiMER provides its bioanalysis service mostly for marine projects, but non-marine ones are also accepted. The platform can provide scientific and technical advices (e.g. choice of analysis methods, optimal use on a supercomputer), carry out bioinformatics and biostatistics analyzes on a project, and intervene in the valorization of scientific results (e.g. participation in the drafting of papers, publishing standardized analysis report following FAIR principles [11]). SeBiMER team can provide expertise in the following fields: community diversity (metabarcoding), metagenomics, genome assembly and annotation, phylogenetics and phylogenomics, genotyping, population genetics and genomics, biostatistics [12, 13].

IV - Software development. SeBiMER can bring to research teams its expertise for the industrialization of methods (use of workflow managers such as Nextflow), the adaptation of existing tools to high performance computing resources, or the creation of new tools adapted to specific data analysis protocols. For the latter, the team provides its expertise to develop softwares following FAIR principles. As examples, SeBiMER is developing tools to analyze eDNA data in the context of Ifremer's Genomics Observatories (SAMBA workflow [14]) and to conduct sequence annotations (ORSON workflow [15]). All softwares developed by SeBiMER are intended to become publicly available over time on our Gitlab and Github profiles [16].

V - Training. The SeBiMER platform provides training services with both scientific and technical goals. For the former, SeBiMER provides trainings to teach users metabarcoding and transcriptomics data analyzes, as well as genomes assembly and annotation. For the latter, SeBiMER helps users to learn Linux operating system, PBS Pro cluster execution environment, software packaging and use with Conda and Singularity, and Nextflow-based workflow use and development. Some of these training sessions are integrated in the Biogenouest and IFB trainings catalogs.

While created as the Ifremer's bioinformatics platform, SeBiMER road map involves its opening to the community outside Ifremer. In this context, on June 2022 the platform became member of Biogenouest, the network of technological core facilities of the Western France in life sciences and the environment. This way, SeBiMER joins ABiMS, Genouest and BiRD bioinformatics platforms [17].

Any requests for a collaboration with SeBiMER can be addressed to corresponding author.

References.

1. <https://www.cnrs.fr/fr/cnrsinfo/pepr-atlasea-plongee-dans-le-genome-du-biotope-marin>
2. <https://www.fairease.eu>
3. <https://www.ifremer.fr/fr/infrastructures-de-recherche/le-supercalculateur-datarmor>
4. <https://github.com/ifremer-bioinformatics/ToolDirectory>
5. <https://github.com/pgdurand/BeeDeeM>
6. <https://gitlab.ifremer.fr/bioinfo/omics-catalog>
7. <https://data.ifremer.fr/SISMER>
8. <https://www.coretrustseal.org>
9. <https://sextant.ifremer.fr/eng/Data/Catalogue>
10. <https://www.france-bioinformatique.fr/en/news/mudis4ls-the-project-for-shared-digital-spaces-for-life-sciences/>
11. https://ifremer-bioinformatics.github.io/SAMBAExampleReport/SAMBA_report.html
12. Clément Bernard, Marie Collard-Paulet, Cyril Noël, *et al.* A time-resolved multi-omics atlas of *Acanthamoeba castellanii* encystment. *Nat Commun.* 13, 4104 (2022). <https://doi.org/10.1038/s41467-022-31832-0>
13. Michèle Gourmelon, Amine Boukerb *et al.* Genomic Diversity of *Campylobacter lari* Group Isolates from Europe and Australia in a One Health Context. *Appl Environ Microbiol.* 2022;88(23):e0136822. <https://doi.org/10.1128/aem.01368-22>
14. <https://github.com/ifremer-bioinformatics/samba>
15. <https://gitlab.ifremer.fr/bioinfo/workflows/orson>
16. <https://gitlab.ifremer.fr/bioinfo> ; <https://github.com/ifremer-bioinformatics/>
17. <https://www.biogenouest.org/plates-formes/>

Session 3: Statistics, machine learning, artificial intelligence and image analysis

DeCovarT: Robust deconvolution of cell mixture in transcriptomic samples by leveraging cross-talk between genes

Bastien CHASSAGNOL^{1,2}, Yufei LUO¹, Gregory NUEL² and Etienne BECHT¹

¹ Les Laboratoires Servier, 50 Rue Carnot, 92150, Suresnes, France

² LPSM (Laboratoire de Probabilités, Statistiques et Modélisation), 4 place Jussieu, 75005, Paris, France

Corresponding author: `bastien.chassagnol@upmc.fr`

Abstract Motivation: *Transcriptomic analyses have contributed greatly to a better understanding of the biological processes involved in the evolution of complex and versatile diseases. However, bulk transcriptomic analyses ignore the contribution of diverse cellular populations to samples heterogeneity. Thus, computational deconvolution methods have been developed to analyse the cellular composition of tissues. However, the performance of these algorithms is limited in distinguishing between cell populations with very similar expression profiles, and we hypothesised that taking into account the covariance between genes could enhance the performance of deconvolution algorithms for closely-related cell populations.*

Results: *We therefore developed a new deconvolution algorithm, DeCovarT, which takes into account the transcriptomic structure of each cell population. To do so, we represented the set of transcriptomic interactions as a multivariate Gaussian distribution, assuming a sparse network structure deduced from the precision matrix returned by the gLasso algorithm. Therefore, we reconstruct the global mixing profile by a generative model, in which we show, under reasonable assumptions, that the law describing the global expression profile conditional on the cell ratios and the purified expression profiles is also a multivariate Gaussian distribution. The maximum likelihood of the associated function, i.e. the cell ratios optimising the probability of observing the observed transcriptomic distribution, is estimated in our paper by a reparametrised, unconstrained version of the Levenberg-Marquardt algorithm. This allows us to estimate more easily an estimator taking into account the unit simplex constraint on the cell ratios while exhibiting more easily a quasi-Gaussian distribution of the estimator.*

Keywords cellular deconvolution, gLasso, generative model, bulk RNA Sequencing, tumor micro environment

1 Introduction

The analysis of the bulk transcriptome using high-throughput sequencing methods provided new insights on the mechanisms involved in the development of diseases. However, such methods ignore the intrinsic cellular heterogeneity of complex biological samples, hampering the identification of the causal drivers of the variability observed between individuals.

Indeed, the cell composition plays a crucial role on the evolution of disease conditions. And classical bulk analyses tend to ignore the intrinsic complexity of biological samples, by averaging measurements over multiple distinct cell populations.

For instance, the tumoral micro environment (TME) encompasses a large variety of cell populations, whose interactions will directly impact the tumoral growth, cancer progression and henceforth the patient outcome. But each cell population displays an unique transcriptomic profile, and even within a cell population the expression patterns can significantly differ, driven by signalling pathways which induce *cell motility* and *cell differentiation* [1].

Not accounting for changes of the cell composition as one of the biological drivers as a confounding signal in downstream analyses, particularly in differential analysis, is likely to result in a loss of *specificity* (genes wrongly identified as differentially expressed, while they only reflect an increase of the cell

population naturally producing them) and *sensibility* (genes expressed by minor cell populations are amenable to be masked by the expression and high-variability of dominant cell populations), which in turn prevents from identifying the true causal drivers of the change of gene dysregulation.

We can set apart two groups of methods for studying cell heterogeneity. Physical methods, such as immunohistochemistry and flow cytometry, can only take profit of a small subset of phenotypic markers to disentangle cell populations, making them burdensome, low-throughput and costly. Single-cell sequencing (scRNASeq) is also a promising avenue, but disassociation of the tissues to isolate single cells prior to sequencing make them inconvenient to analyse deeply intertwined and fibrous tissues.

A whole set of computational methods have been developed for estimating fractions of cell types in bulk admixtures, but they perform badly at discriminating cell types displaying strong phenotypic proximity (e.g., naïve vs. memory B cells, [2]). Indeed, most of them assume that the cellular purified expression profiles are fixed observations, omitting the variability and correlation structure of the genes in reference samples.

In contrast to these approaches, we hypothesised that taking into account the individual variance and pairwise covariance of the genes in the reference transcriptome could enhance the performance of transcriptomic deconvolution methods. We thus introduce *DeCovarT* (Deconvolution using the Transcriptomic Covariance), a new computational and probabilistic approach that may provide less noisy estimates of closely related cell populations.

2 Objective and notations

Similar to most traditional deconvolution models, we assume that the global bulk gene expression is linearly related to the cell purified expression profiles. Precisely, we reconstruct the observed transcriptome from a cellularly-heterogeneous sample by summing the individual contributions of each cell population weighted by its by its corresponding relative frequency in the sample. Formally, let $\mathbf{X} = (x_{gj}) \in \mathcal{M}_{\mathbb{R}^{G \times J}}$ the signature matrix representing the purified transcriptomic profiles of J cell populations and $\mathbf{p} = (p_{ji}) \in]0, 1[^{J \times N}$ the unknown relative proportions of cell populations in N samples, then the linear relation linking the bulk global expression ($\mathbf{y} = (y_{gi}) \in \mathbb{R}_+^{G \times N}$) to the individual cellular expression profiles is represented by the following matricial product: Eq. 1:

$$\mathbf{y} = \mathbf{X} \times \mathbf{p} \tag{1}$$

To be determined, the problem Eq. 1 requires that the number of genes G exceeds the number of cell types and that no purified cellular expression profile can be rewritten as a linear combination of the other cell populations (in other words, that the purified signature profile, \mathbf{X} , is invertible and of full rank J).

However, in presence of technical noise, the strict linearity between the endogenous variable \mathbf{y} and the exogenous variables stored in the design matrix \mathbf{X} does not hold. In that context, the general approach consists of modelling the distortion between the response variable and the corresponding explanatory variables using an unobserved error term. Without further assumption on this disturbance term, the usual approach to minimise the squared error between the mixture expressions predicted by the linear model and the actual observed response is through the ordinary least squares (OLS) approach. Under the Gaussian-Markov assumptions, that we recall in Appendix(Assumptions of the Gaussian linear model), we can show that the OLS estimate returned by this approach is equal to the MLE (maximum likelihood estimate) returned by the model in which we model the noise by an Additive white Gaussian noise Eq. 2:

$$y_{gi} = \sum_{j=1}^J x_{gj} p_{ji} + \epsilon_i, \quad y_{gi} \sim \mathcal{N} \left(\sum_{j=1}^J x_{gj} p_{ji}, \sigma_i^2 \right), \quad \epsilon_i \sim \mathcal{N}(0, \sigma_i^2) \tag{2}$$

In the next section, we consider a new generative model where some of the Gaussian-Markov assumptions are discarded. We also assume for the sake of simplicity that the samples are uncorrelated

(we then drop the index i , the estimation process being similar for each sample: a desired feature for efficient parallel computation).

3 Model

In the DeCovarT approach, we relax the *exogeneity* property (see Theorem **Gauss-Markov** in Appendix *Gaussian-Markov assumptions*), by treating the regressors \mathbf{X} as random variables rather than fixed, deterministic measures. This approach has already been developed using two distinct approaches with the DSection [3] and DeMixt [4] algorithms. But to our knowledge, we are the first to weaken the independence assumption between the observations and to account for the intransitive covariance structure of the transcriptome of reference cell populations. Since most of the variability proceeds from the stochastic nature of the purified expression profiles, we conjecture, for technical identifiability and computational issues, that adding any additional technical error term to the response variable was irrelevant.

Precisely, we assume that the G -dimensional vector \mathbf{x}_j characterising the transcriptomic expression of each cell population follows a multivariate Gaussian distribution: $\mathbf{x}_j \sim \mathcal{N}_G(\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j)$, given in Definition 3.1.

DEFINITION 3.1 (MULTIVARIATE GAUSSIAN DISTRIBUTION). *The multivariate Gaussian distribution of the random vector of size G characterising each purified transcriptomic profile, \mathbf{x}_j , is:*

$$\text{Det}(2\pi\boldsymbol{\Sigma}_j)^{-\frac{1}{2}} \exp\left(-\frac{1}{2}(\mathbf{x}_j - \boldsymbol{\mu}_j)\boldsymbol{\Sigma}_j^{-1}(\mathbf{x}_j - \boldsymbol{\mu}_j)^\top\right)$$

which is parametrised by:

- **a mean vector:** $\boldsymbol{\mu}_j$
- **a positive-definite, full rank covariance matrix.** We use the *gLasso* [5] algorithm to infer it, assuming a largely sparse network structure. We also define $\boldsymbol{\Theta}_j \equiv \boldsymbol{\Sigma}_j^{-1}$ as the precision matrix.

4 Method

To derive the log-likelihood of our model, we supposed first that the *plugged-in* mean and covariance parameters $\zeta_j = (\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j)$ inferred for each purified cell population at the previous step do not differ in our mixed sample (this strong hypothesis is likely to not hold under varying biological heterotypic conditions).

Then, set $\boldsymbol{\zeta} = (\boldsymbol{\mu}, \boldsymbol{\Sigma})$, $\boldsymbol{\mu} = (\boldsymbol{\mu}_j)_{j \in \tilde{J}} \in \mathcal{M}_{G \times J}$, $\boldsymbol{\Sigma} \in \mathcal{M}_{G \times G}$ the supposed known parameters and \mathbf{p} the unknown cellular ratios that parameterise together the conditional distribution $\mathbf{y}|(\boldsymbol{\zeta}, \mathbf{p})$, then the following generative model can be used to rebuild the admixture of cell populations (Fig. 1).

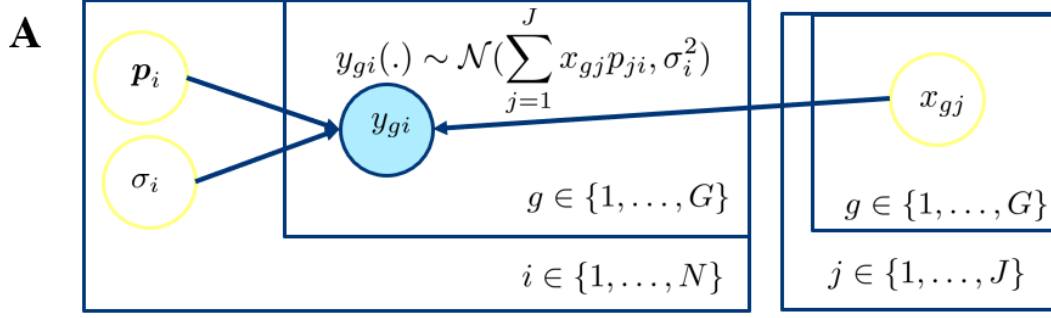
Indeed, keeping the assumption of Independence between covariates (here, the cell populations themselves) and retaining the G stacked linear equations of Eq. 1, the conditional distribution, as a sum of random variables, is the convolution of multivariate pairwise independent Gaussian distributions, from which, using the *affine invariant* property of exponential distributions, it can be shown that it follows the following multivariate Gaussian distribution (Eq. 3):

$$\mathbf{y}|(\boldsymbol{\zeta}, \mathbf{p}) \sim \mathcal{N}_G(\boldsymbol{\mu}\mathbf{p}, \boldsymbol{\Sigma}) \text{ with } \boldsymbol{\mu} = (\boldsymbol{\mu}_j)_{j \in \tilde{J}}, \quad \mathbf{p} = (p_1, \dots, p_J) \text{ and } \boldsymbol{\Sigma} = \sum_{j=1}^J p_j^2 \boldsymbol{\Sigma}_j \quad (3)$$

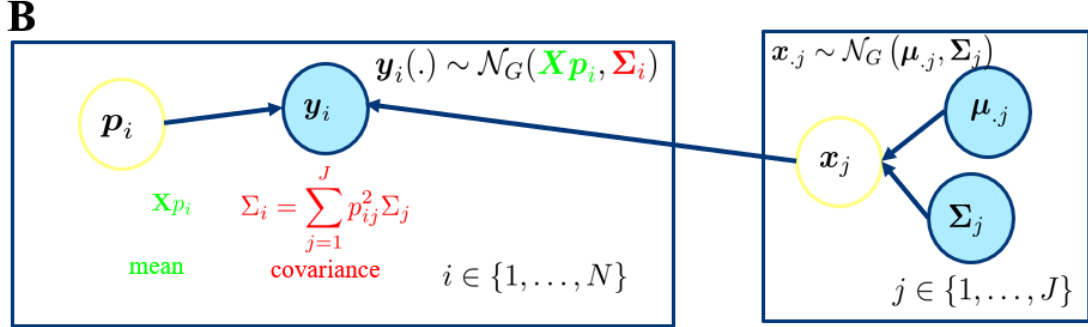
. With this model (Eq. 3), we now plan to estimate the proportions by maximum likelihood and the next sub-sections will detail how to determine the log-likelihood of the distribution function as well as its gradient and hessian, in both an unconstrained and constrained, reparametrised setting.

4.1 Log-likelihood computation

From Eq. 3, we readily compute the log-likelihood of the conditional distribution $\mathbf{y}|(\boldsymbol{\zeta}, \mathbf{p})$ in Eq. 4:



(a) Standard linear model representation.



(b) The generative model used for the DeCovart framework.

Fig. 1. We use the standard representation of a graphical model, with observed states represented as blue-shaded circles and stochastic nodes corresponding to random variables by solid circles. Independent replication over a set of variables is indicated by enclosing the replicated nodes in a rectangle plate, with as index the number of replicates.

$$\ell_{\mathbf{y}|\zeta}(\mathbf{p}) = C + \log \left(\text{Det} \left(\sum_{j=1}^J p_j^2 \Sigma_j \right)^{-1} \right) - \frac{1}{2} (\mathbf{y} - \mathbf{p}\boldsymbol{\mu})^\top \left(\sum_{j=1}^J p_j^2 \Sigma_j \right)^{-1} (\mathbf{y} - \mathbf{p}\boldsymbol{\mu}) \quad (4)$$

with $C = -\frac{G}{2} \log(2\pi)$ a constant.

4.2 First and second-order derivation of the unconstrained DeCovart log-likelihood function

To determine analytically the stationary points of a function, and notably its maximum, we need to determine the roots (the values for a which a function equals zero) of its gradient, in our context, the vector: $\nabla \ell : \mathbb{R}^J \rightarrow \mathbb{R}^J$ evaluated at point $\mathbf{p} = (p_1, \dots, p_J)$ in J -dimensional space. We derive only one component of the gradient vector since the computation is the same for any cell component ratio $p_j, j \in \tilde{J}$ (Eq. 5):

$$\begin{aligned} \frac{\partial \ell_{\mathbf{y}|\zeta}(\mathbf{p})}{\partial p_j} &= \frac{\partial \log(\text{Det}(\boldsymbol{\Theta}))}{\partial p_j} - \frac{1}{2} \left[\frac{\partial (\mathbf{y} - \boldsymbol{\mu}\mathbf{p})^\top}{\partial p_j} \boldsymbol{\Theta} (\mathbf{y} - \boldsymbol{\mu}\mathbf{p}) + (\mathbf{y} - \boldsymbol{\mu}\mathbf{p})^\top \frac{\partial \boldsymbol{\Theta}}{\partial p_j} (\mathbf{y} - \boldsymbol{\mu}\mathbf{p}) + (\mathbf{y} - \boldsymbol{\mu}\mathbf{p})^\top \boldsymbol{\Theta} \frac{\partial (\mathbf{y} - \boldsymbol{\mu}\mathbf{p})}{\partial p_j} \right] \\ &= -\text{Tr} \left(\boldsymbol{\Theta} \frac{\partial \boldsymbol{\Sigma}}{\partial p_j} \right) - \frac{1}{2} \left[-\boldsymbol{\mu}_j^\top \boldsymbol{\Theta} (\mathbf{y} - \boldsymbol{\mu}\mathbf{p}) - (\mathbf{y} - \boldsymbol{\mu}\mathbf{p})^\top \boldsymbol{\Theta} \frac{\partial \boldsymbol{\Sigma}}{\partial p_j} \boldsymbol{\Theta} (\mathbf{y} - \boldsymbol{\mu}\mathbf{p}) - (\mathbf{y} - \boldsymbol{\mu}\mathbf{p})^\top \boldsymbol{\Theta} \boldsymbol{\mu}_j \right] \\ &= -2p_j \text{Tr}(\boldsymbol{\Theta} \Sigma_j) + (\mathbf{y} - \boldsymbol{\mu}\mathbf{p})^\top \boldsymbol{\Theta} \boldsymbol{\mu}_j + p_j (\mathbf{y} - \boldsymbol{\mu}\mathbf{p})^\top \boldsymbol{\Theta} \Sigma_j \boldsymbol{\Theta} (\mathbf{y} - \boldsymbol{\mu}\mathbf{p}) \end{aligned} \quad (5)$$

Since the solution to $\nabla(\ell_{\mathbf{y}|\zeta}(\mathbf{p})) = 0$ is not closed, we instead retrieve the MLE using iterated numerical optimisation methods and some of them requiring a second-order derivation Sec. 4.4. The

second derivative order, corresponding to the Hessian matrix of the log-likelihood (Eq. 4) in $\mathcal{M}_{J \times J}$ is given by:

$$\begin{aligned}
\mathbf{H}_{i,i} &= \frac{\partial^2 \ell}{\partial^2 p_i} = -2 \operatorname{Tr}(\Theta \Sigma_i) + 4p_i^2 \operatorname{Tr}\left((\Theta \Sigma_i)^2\right) - 2p_i(\mathbf{y} - \mu \mathbf{p})^\top \Theta \Sigma_i \Theta \mu_{\cdot i} - \mu_{\cdot i}^\top \Theta \mu_{\cdot i} - \\
&\quad 2p_i(\mathbf{y} - \mu \mathbf{p})^\top \Theta \Sigma_i \Theta \mu_{\cdot i} - (\mathbf{y} - \mu \mathbf{p})^\top \Theta (4p_i^2 \Sigma_i \Theta \Sigma_i - \Sigma_i) \Theta (\mathbf{y} - \mu \mathbf{p}), \quad i \in \tilde{\mathcal{J}} \\
\mathbf{H}_{i,j} &= \frac{\partial^2 \ell}{\partial p_i \partial p_j} = 4p_j p_i \operatorname{Tr}(\Theta \Sigma_j \Theta \Sigma_i) - 2p_i(\mathbf{y} - \mu \mathbf{p})^\top \Theta \Sigma_i \Theta \mu_{\cdot j} - \mu_{\cdot i}^\top \Theta \mu_{\cdot j} - \\
&\quad 2p_j(\mathbf{y} - \mu \mathbf{p})^\top \Theta \Sigma_j \Theta \mu_{\cdot i} - 4p_i p_j (\mathbf{y} - \mu \mathbf{p})^\top \Theta \Sigma_i \Theta \Sigma_j \Theta (\mathbf{y} - \mu \mathbf{p}), \quad (i, j) \in \tilde{\mathcal{J}}^2, i \neq j
\end{aligned} \tag{6}$$

in which the coloured sections pair one by one with the corresponding coloured sections of the gradient, given in Eq. 5. Matrix calculus can largely ease the derivation of complex algebraic expressions, thus we remind in Appendix (*Matrix calculus*) relevant matrix properties and derivations.

4.3 First and second-order derivation of the constrained DeCovarT log-likelihood function

Since \mathbf{p} is constrained (unit simplex constraint on positive cellular ratios, as stated in Eq. 7), we introduce the unconstrained parameter $\boldsymbol{\theta}$ to ease the derivation of the log-likelihood function of the model while accounting for numerical bounds on the inferred proportions:

$$\begin{cases} \sum_{j=1}^J p_j = 1 \\ \forall j \in \tilde{\mathcal{J}} \quad p_j \geq 0 \end{cases} \tag{7}$$

Precisely, we consider the following mapping function: $\boldsymbol{\psi} : \boldsymbol{\theta} \rightarrow \mathbf{p} \mid \boldsymbol{\theta} \in \mathbb{R}^{J-1}, \mathbf{p} \in]0, 1[^J$ (Eq. 4.3) (to ensure the continuity of the mapping function, each ratio must be defined on the opened interval $]0, 1[$):

$$a) \quad \mathbf{p} = \boldsymbol{\psi}(\boldsymbol{\theta}) = \begin{cases} p_j = \frac{e^{\theta_j}}{\sum_{k < J} e^{\theta_k} + 1}, \quad j < J \\ p_J = \frac{1}{\sum_{k < J} e^{\theta_k} + 1} \end{cases} \quad b) \quad \boldsymbol{\theta} = \boldsymbol{\psi}^{-1}(\mathbf{p}) = \left(\ln \left(\frac{p_j}{p_J} \right) \right)_{j \in \{1, \dots, J-1\}}$$

that is a C^2 -diffeomorphism, since $\boldsymbol{\psi}$ is a bijection between \mathbf{p} and $\boldsymbol{\theta}$ twice differentiable. We provide the Jacobian matrix, $\mathbf{J}_{\boldsymbol{\psi}} \in \mathcal{M}_{J \times (J-1)}$ of this vector-valued mapping function in Eq. 8:

$$\mathbf{J}_{i,j} = \frac{\partial p_i}{\partial \theta_j} = \begin{cases} \frac{e^{\theta_i} B_i}{A^2}, & i = j, i < J \\ \frac{-e^{\theta_j} e^{\theta_i}}{A^2}, & i \neq j, i < J \\ \frac{-e^{\theta_j}}{A^2}, & i = J \end{cases} \tag{8}$$

with i indexing vector-valued \mathbf{p} and j indexing the first-order order partial derivatives of the mapping function, $A = \sum_{j' < J} e^{\theta_{j'}} + 1$ the sum over exponential (denominator of the mapping function) and $B = A - e^{\theta_i}$ the same sum, but deprived of the exponential with the same index i as the corresponding component p_i .

We return the symmetric Hessian (the Schwarz's theorem states indeed that switching the second-order partial derivatives of the off-diagonal terms should not impact the result) of the vectorial mapping function $\boldsymbol{\psi}(\boldsymbol{\theta})$ as a third-order tensor (see Sec. *Tensor product*) of level 3 and rank $(J-1)(J-1)J$ in Eq. 9:

$$\frac{\partial^2 p_i}{\partial k \partial j} = \begin{cases} \frac{e^{\theta_i} e^{\theta_l} (-B_i + e^{\theta_i})}{A^3}, (i < J) \wedge ((i \neq j) \oplus (i \neq k)) & (a) \\ \frac{2e^{\theta_i} e^{\theta_j} e^{\theta_k}}{A^3}, (i < J) \wedge (i \neq j \neq k) & (b) \\ \frac{e^{\theta_i} e^{\theta_j} (-A + 2e^{\theta_j})}{A^3}, (i < J) \wedge (j = k \neq i) & (c) \\ \frac{B_i e^{\theta_i} (B_i - e^{\theta_i})}{A^3}, (i < J) \wedge (j = k = i) & (d) \\ \frac{e^{\theta_j} (-A + 2e^{\theta_j})}{A^3}, (i = J) \wedge (j = k) & (e) \\ \frac{2e^{\theta_j} e^{\theta_k}}{A^3}, (i = J) \wedge (j \neq k) & (f) \end{cases} \quad (9)$$

with i indexing \mathbf{p} , j and k respectively indexing the first-order and second-order partial derivatives of the mapping function with respect to $\boldsymbol{\theta}$. In line (a), \oplus refers to the Boolean XOR operator, \wedge to the AND operator and $l = \{j, k\} \setminus i$.

To derive the log-likelihood function in Eq. 5, we reparametrise \mathbf{p} to $\boldsymbol{\theta}$, using a standard *chain rule formula* (see Appendix Calculus notation).

Considering the original log-likelihood function, $\ell :]0, 1[^J \rightarrow \mathbb{R}$, and the mapping function, $\boldsymbol{\psi} : \mathbb{R}^{J-1} \rightarrow]0, 1[^J$ then the differential at the first order is given by Eq. 10:

$$\left[\frac{\partial \ell_{\mathbf{y}|\boldsymbol{\zeta}}}{\partial \theta_j} \right]_{j < J} = \sum_{i=1}^J \frac{\partial \ell_{\mathbf{y}|\boldsymbol{\zeta}}}{\partial p_i} \frac{\partial p_i}{\partial \theta_j} \quad (10)$$

Deriving at the second order once the resulting gradient in Eq. 10 yields the following Hessian matrix, $\mathbf{H}(\ell(\boldsymbol{\theta})) \in \mathcal{M}_{(J-1) \times (J-1)}$, of the constrained log-likelihood function, given by Eq. 11:

$$\left[\frac{\partial^2 \ell_{\mathbf{y}|\boldsymbol{\zeta}}}{\partial \theta_k \partial \theta_j} \right]_{j < J, k < J} = \sum_{i=1}^J \sum_{l=1}^J \left(\frac{\partial p_i}{\partial \theta_j} \frac{\partial^2 \ell_{\mathbf{y}|\boldsymbol{\zeta}}}{\partial p_i \partial p_l} \frac{\partial p_l}{\partial \theta_k} \right) + \sum_{i=1}^J \left(\frac{\partial \ell_{\mathbf{y}|\boldsymbol{\zeta}}}{\partial p_i} \frac{\partial^2 p_i}{\partial \theta_k \partial \theta_j} \right) \quad (d) \quad (11)$$

Using the generalised tensor product, as referred in Appendix (*Tensor product*), we can however ease and compact the numerical implementation of Eq. 11 by avoiding any loops.

4.4 The Levenberg-Marquardt algorithm

When the form of the gradient is non-closed, preventing from analytically retrieving its roots, iterated numerical optimisation algorithms can be used instead as proxys. The second-order descent quadratic approach leverages the local curvature of the function. In that case, the descent direction is given by $d_{\text{Newton-Raphson}} = \frac{\nabla f(\mathbf{p}_i)}{\mathbf{H}_f^{-1}(\mathbf{p}_i)}$, which is generally faster compared to the standard gradient descent based method. However, the estimation of the Hessian matrix is generally burdensome, both in terms of computational resources and theoretical calculus derivation.

Alternatively, the *Levenberg-Marquardt algorithm* [6, 7], bridges the gap between between the steepest descent method and the Newton-Raphson method by inflating the diagonal terms of the Hessian matrix. Precisely, consider the following re-scaling of the Hessian matrix: $\mathbf{H}_f(\mathbf{p}_i)^{\text{LM}} = \mathbf{H}_f(\mathbf{p}_i) + \lambda \mathbf{I}_J$ with \mathbf{I}_J the identity matrix of rank J and λ , a scaling factor regularly updated that controls the contribution of the gradient with respect to the off-diagonal terms of the Hessian matrix. When tending to 0, the descent direction is similar to a second-order one, which is favoured when far from the extremum for its faster convergence pace. On the other hand, when close from the extremum, standard descent gradient with adjusted step size is endorsed, focusing rather on the diagonal terms of the Hessian. Modifying the scaling factor ensures additionally that the Hessian matrix, $\mathbf{H}_f(\mathbf{p}_i^{(q)})$ at step (q) is always positive definite and non-degenerate.

The iterated algorithm stops when convergence criteria are fulfilled, generally when the objective function or the parameters estimated are stabilised. The **marqLevAlg** introduces a new stringent convergence criteria, the relative distance to maximum (RDM), which can be interpreted alternatively as the ratio of the approximation numerical error over the statistical error. Compared to other criteria, it guarantees convergence to an optimum, setting them apart from spurious saddle points [8].

5 Results on simulations

5.1 Simulations

We asserted numerically the relevance of accounting the correlation between expressed transcripts, we designed a simple toy example with two genes and two cell proportions. Hence, with the simplex constraint (Eq. 7), we only have to estimate one free unconstrained parameter, θ_1 , and then use the mapping function Eq. 4.3 to recover back the ratios.

We rebuilt the bulk mixture, $\mathbf{y} \in \mathcal{M}_{\mathbb{R}^{G \times N}}$, for a set of artificial samples $N = 500$ with the following generative model:

- We test two levels of cellular entropy: one with balanced ($\mathbf{p} = (p_1, p_2 = 1 - p_1) = (\frac{1}{2}, \frac{1}{2})$) cell populations and one with highly unbalanced cell populations: $\mathbf{p} = (0.95, 0.05)$.
- Then, each cellular purified transcriptomic profile is drawn from a multivariate Gaussian distribution. We compared two scenarios, one with close centroids and one with far centroids, respectively $\mu_{.1} = (20, 22), \mu_{.2} = (22, 20)$ and $\mu_{.1} = (20, 40), \mu_{.2} = (40, 20)$. For the covariance matrix, $\Sigma \in \mathcal{M}_{2 \times 2}$ covariance matrix, we consider equibalanced variances for the marginal distribution of each gene, $\text{Diag}(\Sigma_1) = \text{Diag}(\Sigma_2) = \mathbf{I}_2$, playing only on the level of correlation between gene 1 and gene 2, $\text{Cov}[x_{1,2}] = \text{Cov}[x_{2,1}]$, testing all paired ranges in population 1 and 2 from -0.8 to 0.8, by 0.2 step size.
- As stated in 1, we assume that the bulk mixture, \mathbf{y}_i could be directly reconstructed by summing up the individual cellular contributions weighted by their abundance, without additional noise.

To infer back the cellular ratios for each of these samples, we consider the plugged-in parameters of the individual mean and covariance expression profiles known, and we optimise the vectorial parameter \mathbf{p} that maximises the log-likelihood distribution stated in 4, using the implementation of the Levenbergh-Marquard algorithm supplied in R package `marqLevAlg`. We compared this approach as a standard non-negative least squares approach (NNLS) estimated through the *Lawson Hanson algorithm* [9] which aims at reducing the quadratic error between the reconstituted bulk expressions predicted by linear regression while enforcing the non-negativity constraint on the ratios.

5.2 Results

The overlap between two probability distributions depends on the proximity of their centroids, the marginal variability of the individual random variables and the level of correlation. Setting all the other varying parameters that contribute to the level of overlap, we highlight numerically that the degree of correlation between two genes (Fig. 2) strongly impacts the overall performance.

More precisely, the level of overlap between two cell population distributions is an excellent predictor of the quality of the estimation: the less the distributions of two cell populations overlap, the better the specificity of the estimation. We also show on this small example that not only the distance between centroids (i.e. differences in gene expression means) but also the intrinsic covariance structure allow us to unambiguously characterise each cell population. Indeed, we systematically obtain better results than with a traditional constrained linear approach (LSEI) which assumes independent observations on highly overlapping populations.

Thus, unlike the traditional feature selection approach, which advocates minimising correlation between populations and discarding highly correlated genes, as in the AutoGeneS approach [11], we are able to take advantage of a new set of genes with closely related expression profiles but distinct co-expression patterns. We therefore believe that the greater flexibility of our deconvolution algorithm will improve the currently poor cellular resolution of deconvolution methods, with an increased ability to discriminate between highly correlated cell populations.

6 Perspectives

The new deconvolution algorithm that we implemented, DeCovart, is the first one based on a multivariate generative model while complying explicitly the simplex constraint. Hence, it provides

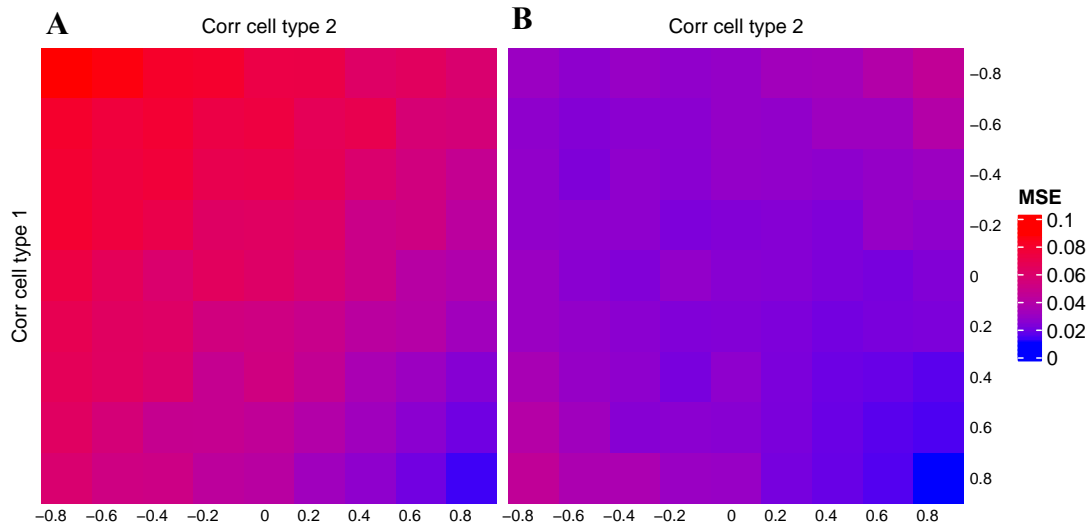


Fig. 2. We display the average mean squared error (MSE) of the estimated cellular ratios comparing a Least Squares with Equality and Inequality Constraints (LSEI) approach, as implemented in the deconRNASeq algorithm [10], to the left, with our newly implemented DeCovarT algorithm, using the robust Marquardt-Levenberg optimisation supplied with the R package `marqLevAlg` (to the right). We use the `ComplexHeatmap` package to draw both Heatmap representations on the same scale.

a strong basis to further derive theoretical confidence bands and generate statistical tests to assert whether a cell population is missing or not, or if the proportion of a cell population differs between two distinct biological conditions.

However, we still need to assert its performance in a real-world experience, by benchmarking it comprehensively against blood and tumoral samples in biological experiments that display sample bulk mixture paired with cytometric measures. We could notably take profit of the huge compendium generated by the *Kassandra* project to that purpose. Indeed, the *Kassandra* database collects 9,404 samples annotated into 38 blood populations. Additionally, the performance of the *Kassandra* algorithm was validated on a total of $N = 517$ samples in 6 public datasets with both flow cytometry and RNA-seq analysis, and benchmarked against 8 different standard deconvolution algorithms: 5 reference profile deconvolution algorithms EPIC [12], CIBERSORT [13], CIBERSORTx [14], *quantIseq* [15] and ABIS [16], and 3 marker-based deconvolution algorithms MCPcounter [17], *xCell* [18] and *Scaden* [19].

Finally, intensive work has still to be done to refine the plug-in parameters provided to our deconvolution optimisation algorithm. We use the *gLasso* [5] algorithm to infer a sparse network structure, encoded by the precision matrix, to describe the transcriptomic interactions occurring within a reference cell population. However, as any penalty regularisation approach, *gLasso* is hampered by *parameter shrinkage*, entailing in practice that the non-null partial correlations are generally underestimated. A way to circumvent this problem is to only use the *support* (the non-null inputs) output of the *gLasso* to fine-tune the estimation of the network using a standard maximum likelihood strategy. To do so, we would integrate the topological constraints induced by the null inputs of the precision matrix to learn a directed Gaussian Graphical Model (GGM).

We detail additional theoretical results, such as matrix calculus and notations, asymptotic distribution of the estimator and associated statistical tests as well as additional simulation results in the vignettes of the Github account of the project: [DeCovarT](#).

References

- [1] Shoemaker, Jason E. and Lopes, Tiago JS and others. CTen: A Web-Based Platform for Identifying Enriched Cell Types from Heterogeneous Microarray Data. *BMC Genomics*, 2012.
- [2] Gregor Sturm, Francesca Finotello, Florent Petitprez, Jitao David Zhang, Jan Baumbach, Wolf H. Fridman, Markus List, and Tatsiana Aneichyk. Comprehensive evaluation of transcriptome-based cell-type quantification methods for immuno-oncology. *Bioinformatics (Oxford, England)*, 2019.
- [3] Erkkilä, Timo and Lehmusvaara, Saara and others. Probabilistic Analysis of Gene Expression Measurements from Heterogeneous Tissues. *Bioinformatics*, 2010.
- [4] Wang, Zeya and Cao, Shaolong and others. Transcriptome Deconvolution of Heterogeneous Tumor Samples with Immune Infiltration. *iScience*, 2018.
- [5] Mazumder, Rahul and Hastie, Trevor. The Graphical Lasso: New Insights and Alternatives. *Electronic Journal of Statistics*, 2011.
- [6] Levenberg, Kenneth. A method for the solution of certain non-linear problems in least squares. *Quarterly of Applied Mathematics*, 1944.
- [7] Marquardt, Donald W. An Algorithm for Least-Squares Estimation of Nonlinear Parameters. *Journal of the Society for Industrial and Applied Mathematics*, 1963.
- [8] Mélanie Prague, Daniel Commenges, Jérémie Guedj, Julia Drylewicz, and Rodolphe Thiébaud. NIMROD: A program for inference via a normal approximation of the posterior in models with random effects based on ordinary differential equations. *Computer Methods and Programs in Biomedicine*, 2013.
- [9] Karen H. Haskell and Richard J. Hanson. An algorithm for linear least squares problems with equality and nonnegativity constraints. *Mathematical Programming*, 1981.
- [10] Gong, Ting and Szustakowski, Joseph D. DeconRNASeq: A Statistical Framework for Deconvolution of Heterogeneous Tissue Samples Based on mRNA-Seq Data. *Bioinformatics (Oxford, England)*, 2013.
- [11] Aliee, Hananeh and Theis, Fabian J. AutoGeneS: Automatic Gene Selection Using Multi-Objective Optimization for RNA-seq Deconvolution. *Cell Systems*, 2021.
- [12] Racle, Julien and de Jonge, Kaat and others. Simultaneous Enumeration of Cancer and Immune Cell Types from Bulk Tumor Gene Expression Data. *eLife*, 2017.
- [13] Newman, Aaron and Liu, Chih and others. Robust Enumeration of Cell Subsets from Tissue Expression Profiles. *Nature methods*, 2015.
- [14] Aaron M. Newman, Chloé B. Steen, Chih Long Liu, Andrew J. Gentles, Aadel A. Chaudhuri, Florian Scherer, Michael S. Khodadoust, Mohammad S. Esfahani, Bogdan A. Luca, David Steiner, Maximilian Diehn, and Ash A. Alizadeh. Determining cell type abundance and expression from bulk tissues with digital cytometry. *Nature Biotechnology*, 2019.
- [15] Finotello, Francesca and Mayer, Clemens and others. Molecular and Pharmacological Modulators of the Tumor Immune Contexture Revealed by Deconvolution of RNA-seq Data. *Genome Medicine*, 2019.
- [16] Gianni Monaco, Bernett Lee, Weili Xu, Seri Mustafah, You Yi Hwang, Christophe Carré, Nicolas Burdin, Lucian Visan, Michele Ceccarelli, Michael Poidinger, Alfred Zippelius, João Pedro de Magalhães, and Anis Larbi. RNA-Seq Signatures Normalized by mRNA Abundance Allow Absolute Deconvolution of Human Immune Cell Types. *Cell Reports*, 2019.
- [17] Etienne Becht, Nicolas A. Giraldo, Laetitia Lacroix, Bénédicte Buttard, Nabila Elarouci, Florent Petitprez, Janick Selves, Pierre Laurent-Puig, Catherine Sautès-Fridman, Wolf H. Fridman, and Aurélien de Reyniès. Estimating the population abundance of tissue-infiltrating immune and stromal cell populations using gene expression. *Genome Biology*, 2016.
- [18] Aran, Dvir and Hu, Zicheng and others. xCell: Digitally Portraying the Tissue Cellular Heterogeneity Landscape. *Genome Biology*, 2017.
- [19] Kevin Menden, Mohamed Marouf, Sergio Oller, Anupriya Dalmia, Daniel Sumner Magruder, Karin Kloiber, Peter Heutink, and Stefan Bonn. Deep learning-based cell composition analysis from tissue expression profiles. *Science Advances*, 2020.

Assessing goats' fecal avoidance using image analysis based monitoring

Mathieu BONNEAU¹, Xavier GODARS² and Jean-Christophe BAMBOU¹

¹ UR ASSET, Domaine Duclos, 97170, Petit-Bourg, Guadeloupe

² UE PTEA, Domaine Duclos, 97170, Petit-Bourg, Guadeloupe

Corresponding author: mathieu.bonneau@inrae.fr

*Reference paper: Bonneau et al. (2022) Assessing goats' fecal avoidance using image analysis based monitoring. *Frontiers in Animal Science*. <https://www.frontiersin.org/articles/10.3389/fanim.2022.835516/full>*

1 Introduction

Recent advances in computer vision (CV) open new perspectives for monitoring animal behavior using CCTV cameras. The process of monitoring behavior through videos is decomposed into several steps: (i) film the animal during the monitoring period, (ii) detect the animal on the videos, (iii) compute any behavioral variables. Behavioral variables could be computed directly from the detection, it is for example the case when one wants to estimate animal location or speed. In other cases, extra treatments are necessary, for example to determine the animal pose or activity, which requires using other computer vision methods. Although major progresses in CV have been made recently, there is a lack of a generic tool that can be used by most scientists, without expertises in CV. This work is an illustration of a CV application to study animal behavior. The level of infestation of goats that are raised outside can generally be described using a log-normal distribution: few animals are highly infested by parasites, while most animals are not, or little infested. Many variables can explain this distribution, such as genetic, parasitic history, nutrition or behavior. In this article we explored the potential of animals to prevent infestation by avoiding feces.

2 Material and Methods

Four male Creole goats were allowed to graze a square pasture of 12m by 12m during 14 days: 7 consecutive days during the first try and 7 more days during the second try, more than two months later. Before the animal entered the pasture, it was contaminated with infested feces, dropped regularly in two rectangular plots located inside the pasture. The aim of the study was to use the camera to estimate and compare the time spent from each animal inside the two contaminated areas.

Before each try, all animals were drenched and worm free. They get infested while grazing during the try, and kept inside after, where no more contamination were possible. Their level of infestation were estimated using fecal eggs count.

Their were recorded using time-lapse camera, taking one picture every 20s, from 6:30am to 6pm.

Monitoring consists into four steps: (i) film animals at pasture, (ii) detect the animal automatically and (iii) identify the detected animals, (iv) estimate the animals' location.

For animal detection, we used Yolo, trained on 3,820 images. For animal identification, we trained a classification CNN, named resNet-50. The CNN was trained in order to identify the animals detected by Yolo. The animals were previously choose for their different color: black, white, red and brown. Each color was then a class of the CNN. 12,236 images were labeled for training and testing (70% vs 30%).

3 Results

First, 87.95% of the goats were detected and they were identified with a precision of 94.8% and a sensitivity of 94.9%. Animals expressed various avoidance level, but generally increases during the week, with fewer feed available. Interestingly, the more animals get infested during the first try, the more they increased their avoidance level during the second try.

Multivariate Analysis of RNA Chemistry Marks Uncovers Epitranscriptomics-Based Biomarker Signature for Adult Diffuse Glioma Diagnostics

Sébastien RELIER¹, Amandine AMALRIC^{1,2}, Aurore ATTINA², Izouline Blaise KOUMARE^{3,4}, Valérie RIGAU⁵, Fanny BUREL VANDENBOS⁶, Denys FONTAINE⁶, Marc BARONCINI⁷, Jean-Philippe HUGNOT¹, Hugues DUFFAU^{1,3}, Luc BAUCHET^{1,3}, Christophe HIRTZ², Eric RIVALS⁸ and Alexandre DAVID¹

¹ IGF, Univ. Montpellier, CNRS, INSERM, Montpellier, France.

² IRMB-PPC, INM, Univ Montpellier, CHU Montpellier, INSERM CNRS, Montpellier 34295, France

³ Neurosurgery Department, Montpellier University Medical Center, Montpellier, Hérault 34295, France

⁴ Neurosurgery Department, CHU Gabriel Toure, Bamako, Mali

⁵ Department of Pathology and Oncobiology, Montpellier University Medical Center, Montpellier, Hérault 34295, France

⁶ Central Laboratory of Pathology, Univ. Côte d'Azur, CHU Nice, CNRS, INSERM, Nice, Alpes-Maritimes 06000, France

⁷ Neurosurgery Department, CHU Lille, Univ. of Lille, Lille, Nord 59037, France

⁸ LIRMM, Université Montpellier, CNRS, F-34095, Montpellier, France

Corresponding author: eric.rivals@lirmm.fr

Reference paper: Relier *et al.* (2022). Multivariate Analysis of RNA Chemistry Marks Uncovers Epitranscriptomics-Based Biomarker Signature for Adult Diffuse Glioma Diagnostics. *Analytical Chemistry*. <http://dx.doi.org/10.1021/acs.analchem.2c01526>

One of the main challenges in cancer management relates to the discovery of reliable biomarkers, which could guide decision-making and predict treatment outcome. In particular, the rise and democratization of high-throughput molecular profiling technologies bolstered the discovery of “biomarker signatures” that could maximize the prediction performance. Such an approach was largely employed from diverse OMICs data (i.e., genomics, transcriptomics, proteomics, metabolomics) but not from epitranscriptomics, which encompasses more than 100 biochemical modifications driving the post-transcriptional fate of RNA: stability, splicing, storage, and translation. We and others have studied chemical marks in isolation and associated them with cancer evolution, adaptation, as well as the response to conventional therapy. In this study, we have designed a unique pipeline combining multiplex analysis of the epitranscriptomic landscape by high-performance liquid chromatography coupled to tandem mass spectrometry with statistical multivariate analysis and machine learning approaches in order to identify biomarker signatures that could guide precision medicine and improve disease diagnosis. We applied this approach to analyze a cohort of adult diffuse glioma patients and demonstrate the existence of an “epitranscriptomics-based signature” that permits glioma grades to be discriminated and predicted with unmet accuracy. This study demonstrates that epitranscriptomics (co)evolves along cancer progression and opens new prospects in the field of omics molecular profiling and personalized medicine.

This work has been published in 2022 in ACS ANALYTICAL CHEMISTRY [1].

Acknowledgements

The project was supported by the Occitanie Region/FEDER (PPri, SMART project), by the Institut National du Cancer (INCa) DA N°2020-116 (AAP PLBIO 2020 - DAVID), *Ligue Contre le Cancer*, SIRIC Montpellier Cancer (INCa-DGOS-Inserm 6045), CHU Montpellier, ARTC Sud, Des Etoiles dans la Mer, and Cancéropole GSO.

References

- [1] S. Relier, A. Amalric, A. Attina, I.B. Koumare, V. Rigau, F. Burel Vandebos, D. Fontaine, M. Baroncini, J.P. Hugnot, H. Duffau, L. Bauchet, C. Hirtz, E. Rivals, and A. David. Multivariate analysis of rna chemistry marks uncovers epitranscriptomics-based biomarker signature for adult diffuse glioma diagnostics. *Analytical Chemistry*, Aug 2022.

LEAF: a machine learning approach to predict effector proteins in *Candidatus* Phytoplasma

Giulia CALIA^{1,2}, Paola PORRACCILO^{3,4}, Djampa KOZLOWSKI^{3,4}, Etienne G.J. DANCHIN⁴, Claudio DONATI², Hannes SCHULER^{1,5}, Mirko MOSER², Alessandro CESTARO², Silvia BOTTINI³

¹ Free University of Bolzano, Faculty of Science and Technology-Faculty of Agricultural, Piazza Università 1, 39100, Bolzano, Italy

² Fondazione Edmund Mach, Research and Innovation Centre, Via Edmund Mach 1, 38098, San Michele all'Adige, Italy

³ Université Côte d'Azur, Center of Modeling, Simulation and Interactions, Route de Saint Antoine de Ginestiere 151, 06200, Nice, France

⁴ INRAE, Université Côte d'Azur, CNRS, Institut Sophia Agrobiotech, Route des Chappes 400, 06903, Sophia-Antipolis, France

⁵ Free University of Bolzano, Competence Centre for Plant Health, Piazza Università 5, 39100, Bolzano, Italy

Corresponding Author: Giulia.Calia@student.unibz.it

Abstract

*Phytoplasmas are plant-parasite, insect vector-borne, bacteria causing enormous economic losses around the globe, infecting many crop plants. Because of their parasitic nature, they cannot be cultivated in-vitro and can be stably maintained only in plants. This has hampered the understanding of their biology and their interaction with both plant and insect hosts. Phytoplasmas encode specific pathogenicity factors called effector proteins. These proteins interact with host protein targets and interfere with the host metabolism causing different levels of symptoms like the abnormal proliferation of shoots, yellowing of the leaves, and dwarfism, to mention a few. Despite being central in the relationship between these pathogens and host-plants, identification of effector proteins remains challenging. Indeed, among the 49 phytoplasmas with available sequencing information only 5 effectors are characterized, namely SAP05, SAP11, SAP54, PHYLI, and TENGU. Here we present LEAF, a machine-learning approach based on a random forest model to predict effector proteins in phytoplasmas. LEAF is trained on a total of 479 proteins from different phytoplasma species. We combined features extracted from the literature description of small secreted proteins with the physicochemical characteristics of protein sequences by using a novel software called MOnSTER. Comparing the resulting predictions on 6 proteomes of phytoplasmas with those of standard and recent methods for effector prediction in bacteria, we proved that LEAF has competitive predictive power of known and putative new effector proteins in the *Candidatus* phytoplasma genus.*

Keywords

Machine Learning; Prediction; Effectors; Phytoplasma; Random-Forest

1 Introduction

LEAF: a machine learning approach to predict effector proteins in *Candidatus Phytoplasma*

Giulia Calia^{1,2}, Paola Porracciolo^{3,4}, Djampa Kozłowski^{3,4}, Etienne G.J. Danchin⁴, Claudio Donati², Hannes Schuler^{1,5}, Mirko Moser², Alessandro Cestaro², Silvia Bottini³

¹ Free University of Bolzano, Faculty of Science and Technology-Faculty of Agricultural, Piazza Università 1, 39100, Bolzano, Italy

² Fondazione Edmund Mach, Research and Innovation Centre, Via Edmund Mach 1, 38098, San Michele all'Adige, Italy

³ Université Côte d'Azur, Center of Modeling, Simulation and Interactions, Route de Saint Antoine de Ginestiere 151, 06200, Nice, France

⁴ INRAE, Université Côte d'Azur, CNRS, Institut Sophia Agrobiotech, Route des Chappes 400, 06903, Sophia-Antipolis, France

⁵ Free University of Bolzano, Competence Centre for Plant Health, Piazza Università 5, 39100, Bolzano, Italy

Giulia.Calia@student.unibz.it

Phytoplasmas are plant-pathogenic, phloem-restricted, bacteria assigned to the class *Mollicutes*. They are cell wall-less pleomorphic bacteria of 0,2-0,8 μm in size that infect both ornamentals and crop plants, causing huge economic losses per year, worldwide [1,2,3]. Only recently it was discovered that phytoplasmas encode pathogenicity factors strongly linked with the progress of the disease. These factors are described as small secreted proteins able to interfere with the host metabolism and are defined as effector proteins. To date, the best-characterized effector proteins in phytoplasmas are: TENGU, which causes dwarfism and altered flower structures [4]; SAP05 which interferes with plant vegetative growth [5]; SAP11 which causes abnormal proliferation of young shoots and changes in leaves shape [6,7]; SAP54/PHYL1 two homologous effectors that cause phyllody symptoms (flowers develop into leaf-like flowers) [8,9].

The impossibility to cultivate phytoplasmas under axenic conditions because of their parasite behavior, hinder the experimental studies focused on the identification of effector proteins. Similarly, *in silico* identification of effector proteins exclusively based on the overall sequence similarities is inefficient due to their protein sequence variability. Nowadays, current methods for effector identification in phytoplasmas are only based on the presence of signal peptide, but it is important to consider that not all the secreted proteins are effectors and that some effector proteins are secreted by nonclassical pathways [10]. Moreover, it is shown that in gram + bacteria, ancestors of phytoplasmas, the h-region of signal peptides is longer than usual, making them more similar to transmembrane regions and undetectable by software specifically designed for signal peptide prediction [11].

The identification of further effector proteins would contribute to the understanding of the biological mechanisms enhancing the disease development, providing new knowledge for the development of novel intervention strategies for the pest management. A recent attempt to develop a tool to predict effectors based on learning models is Deepredef, a convolutional neural network trained on bacteria sequences (gram+ and gram -) [12]. However, there is an urgent need for further novel approaches to improve the prediction of effector proteins. Lately the application of learning models to prediction tasks in biology has shown a great potential [13]. Hence the aim of our work is to build a computational method employing a learning model to efficiently predict effector proteins. Due to the fragmented understanding of phytoplasmas effectors, we first performed an extensive literature mining to identify characteristics to better describe these proteins. We found that sequence length and signal peptide are the most used features for effector prediction but the latter suffers from the above-mentioned problems. To compensate for this we introduced, as features, the

predictions of the transmembrane domain. Since it is shown that some effector proteins exhibit intrinsically disordered regions, we also introduced this as a feature. Then, to complete the set of features, we focused on functional protein motifs and also developed the MOnSTER software. Briefly, by studying the physicochemical characteristics of a set of sequences, MOnSTER creates clusters of short sequence motifs found in protein sequences (CLUMPs). The assumption of MOnSTER is that different amino acids can share similar physicochemical traits and that implies similar properties of these motif sequences. Therefore, it provides a measure of shared motif characteristics in contrast with the sole motif abundance used by other available methods [14, 15, 16].

Afterwards we fed LEAF, a collection of random forest models, with a training dataset consisting of both validated and automatically annotated effector proteins, as positive class, and proteins known to be functionally different from effector proteins having curated annotations, as negative class. Among the several supervised learning algorithms, we choose the random forest because of the small and slightly unbalanced set of known examples at our disposal, and its robustness to outliers, feature correlation, and mixed feature types. We collected, in total, 30 features calculated on 479 proteins and we used different feature combinations to develop three random forest models. Each random forest uses a 5-fold Cross Validation approach for training and testing itself on unseen data, thus using the entire starting dataset. The model takes advantage of the bagging ensemble method to perform the classification. The idea behind the bagging method is that the final prediction benefits from the combination of the predictions of different models. Thus the results of the three instances of the model are combined in a consensus prediction. From the comparison with other current methods for effectors prediction, we show that LEAF outperformed them and its application reached comparable but more exhaustive performances when applied to 6 proteomes of *Candidatus phytoplasma*.

2 Document Structure

2.1 Extended methods

For the training dataset selection and curation, we used the Uniprot database (release 2022_01) starting with choosing proteins uniquely belonging to the phytoplasmas TAXID 33926. Proteins annotated as Effector, SAP05, SAP11, SAP54, TENGU, or PHYL1/phyllody, are selected to be the positive class. In total, the training set contains 184 effector protein sequences. Conversely, proteins with a manually reviewed annotation that is not referring to effectors, are included in the negative class, for a total of 295 non-effector proteins. No putative effectors or protein fragments are included in the training dataset.

To extract characteristic features of effectors, several publicly available software are used (SignalP4.1 [17], TMHMM [18], MobiDB-lite [19], Prosite [20]), except for MOnSTER that we developed in the present study. MOnSTER takes as inputs a list of protein motifs and two fasta files containing positive and negative class sequences. It measures 13 physicochemical properties of motifs (e.g. pH, hydrophathy, amino acids dimension, and composition), and creates clusters of motifs (CLUMPs). MOnSTER uses a discriminative approach exploiting the presence of a predicted list of motifs and measuring their enrichment in the positive class. A new scoring function is introduced with this software, called MOnSTER-score, assigned to each detected CLUMP. This score is an integration of three different scores, the former focused on giving more importance to CLUMPs' amino acid composition, preferring motif sequences present in the positive class, and the latest two based on the Jaccard index modified to prefer occurrences of motifs in the positive class and the number of sequences in the positive class presenting motifs in the considered CLUMP. The higher the MOnSTER-score the more the CLUMP is specific for the positive class.

The final selected features include (i) sequence length; (ii) signal peptide (using the D-score of SignalP 4.1 software); (iii) transmembrane domain (TM) information, more specifically, (iii.1) number of predicted TMs, (iii.2) expected number of aminoacids in TMs, (iii.3) expected amino acids in the first 60 positions of TM helices, (iii.4) probability that the N-term is on the inner side of the membrane, (iii.5) warning for signal peptide N-term (if $\text{iii.3} > 10$); (iv) length of intrinsically disordered regions, when present; (v) occurrences of functional protein motifs of positive class; (vi) occurrence of selected CLUMPs by MOnSTER-score $>$ mean CLUMPs scores; (vii) four features each representing consecutive 25% of a protein sequence and having, as feature-value, the occurrence of motifs belonging to previous selected CLUMPs. All the software cited before are used with the default parameters except for SignalP 4.1 which is configured as in the work of

2.2 Main findings

The average performances of the three LEAF models reached an accuracy of 97%, a weighted F-measure of 97,3%, a precision of 96,9%, and a recall of 96,4%.

To evaluate the performances of our model we also compared its performances with the only two other methods available: SignalP and Deepredef, on the training set. Performances derived from SignalP4.1 reached an accuracy of 92,6% (weighted F-measure of 92,6%), a precision of 92,5%, and a recall of 88%. The lower performance of this method compared to LEAF is a reflection of the presence of signal peptides in proteins that are not effectors and effector proteins for which the signal peptide is not predicted by SignalP4.1. The lowest performances are reached by Deepredef for which the accuracy and the F-measure, 16,7% and 39,5%, respectively, show an important drop compared with the other two methods resulting in 72,3% of information loss. Here both positive and negative classes are misclassified by the model emphasizing the peculiarities of phytoplasmas effector proteins. Importantly, LEAF outperformed both the above-mentioned software that showed lower sensitivity compared to it (Table 1).

To show the performances of LEAF on protein sequences of interest, we selected the best model between the 5-fold Cross-Validation for each of the three models and used them to predict effector proteins in 6 phytoplasma proteomes never seen by the models, namely: *Candidatus* phytoplasma oryzae, strains Mb1a1 (UP000249343), and NGS-S10 (UP000070069), *Candidatus* phytoplasma phoenicium, strains ChiP (UP000238672), and SA213 (UP000037086), *Candidatus* phytoplasma pruni, strain CX (UP000037386), and Maize bushy stunt phytoplasma (CP015149, included in the 16S group of *Candidatus* phytoplasma asteris).

Using the ensemble method for the predictions, LEAF classified, on average, 17% of effector proteins in the overall phytoplasma proteomes.

The application of SignalP4.1 and Deepredef on the same proteomes have resulted, on average, in 26%, of effector proteins predicted by SignalP4.1 and on average 66% effector proteins for Deepredef (Table 1). Altogether these results suggest the validity of our novel method and rule out Deepredef from further analysis on phytoplasmas effectors.

	Accuracy (%)	F-measure (%)	Precision (%)	Recall (%)	Effectors in 6 Ca.P. proteomes(%)
LEAF	97,4 (±1,4%)	97,3 (±1,5%)	96,9 (±1,9%)	96,4 (±2,7%)	17 (±3.9%)
SignalP4.1	92,6	92,6	92,5	88	26 (±3%)
Deepredef	16,7	39,5	5	6,5	66 (±2,8%)

Table 1. Summary of different tools performances on the training set and averaged proportion of predicted effectors in 6 proteomes of phytoplasmas.

To further investigate the characteristics of putative effector proteins predicted by SignalP4.1 or LEAF, we performed a compositional analysis on annotations of the totality of the predictions (Fig. 2). In particular, the annotations of predicted proteins by SignalP4.1 comprise not only effector proteins but also ribosomal proteins, permeases, many uncharacterized proteins, and almost no putative effector proteins (Fig. 2b and c). On the other hand, putative effector proteins are differently abundant in the annotation of effectors by LEAF, together with hypothetical proteins, FtsH proteases, and AAA+ ATPases (Fig. 2a and c). These last categories are very interesting, especially considering the higher number of FtsH proteases predicted by LEAF (Fig. 2c). In fact, it has recently been demonstrated that polymorphisms in FtsH proteases and AAA+ ATPases proteins can be related to virulence in Apple Proliferation phytoplasma strains causing mild to severe symptoms in apple plants, linking these proteins with pathogenicity characteristics [21]. These

findings highlight the potentiality of LEAF in the refinement of effector proteins prediction with reliable actual annotations. The fact that all the mentioned effector-related annotations detected by LEAF fall in the range of 100-95% of positive class-probability predictions suggest an efficient skimming of the putative effectors to be experimentally validated.

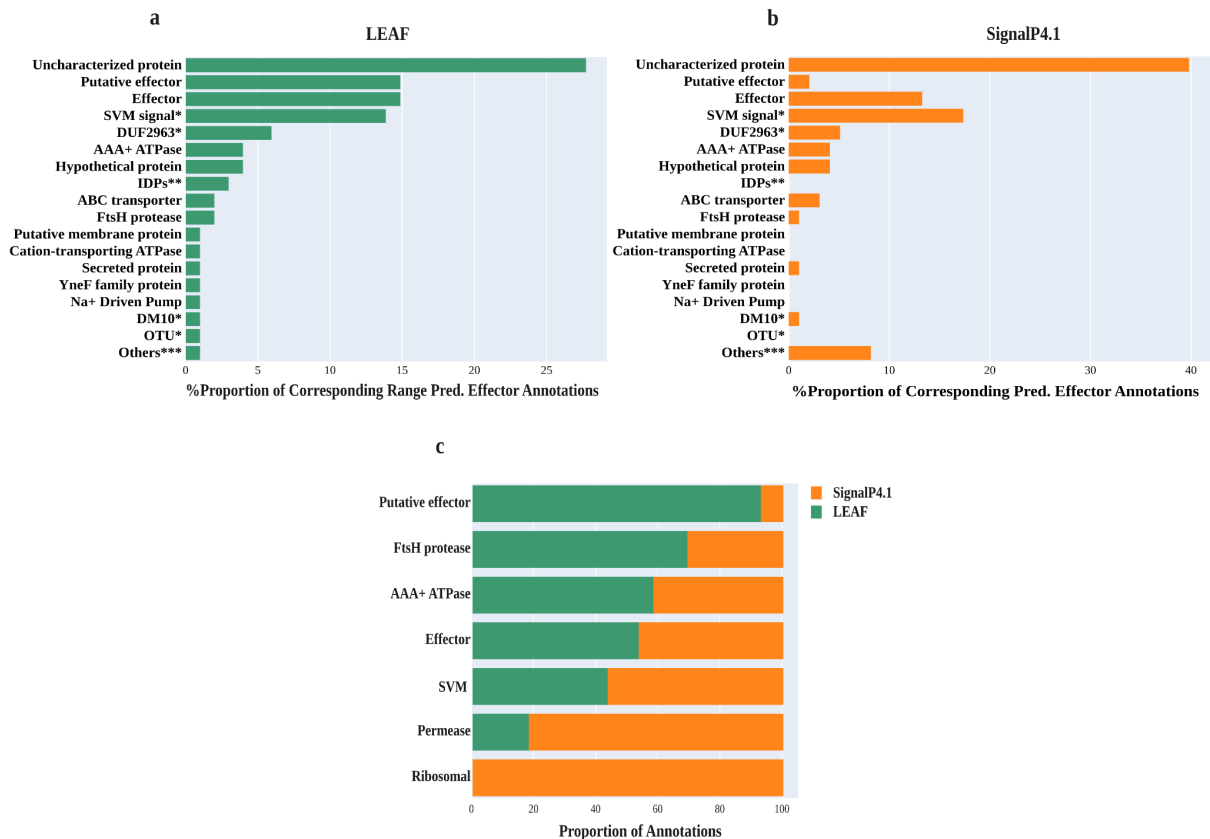


Fig 2. All the figures consider the overall annotations from the 6 phytoplasmas proteomes. a. Representation of annotations proportion for effectors having prediction of positive class-probabilities $\geq 95\%$ assigned by LEAF. b. Due to different metrics for effector predictions, a comparison with a. is made by considering the annotations for the same number of predicted effectors, ranked by D-score of SignalP4.1. c. Representation of different abundances in annotations for predicted effectors between the overall set of annotations from LEAF and SignalP4.1. *domain-containing protein, **Immunodominant membrane proteins, ***Chromosome segregation ATPase, Prolyl oligopeptidase family protein, Acyltransferase family protein, Lipoprotein, Solute-binding protein family.

2.3 Conclusions and further directions

The development of a novel effector predictor in phytoplasmas is aiming to compensate for the lack of reliable methods for this scope taking into account the biological complexity of these specific and elusive proteins. We demonstrated that by combining efficient feature selection of peculiar traits of effector proteins in the Candidatus Phytoplasma genus, and supervised learning strategy as random forest it is possible to outperform standard and recent methods in the prediction of phytoplasmas effector proteins. LEAF, indeed, provides a more accurate list of possible candidate effectors that, thanks to the positive class-probability value assigned to each prediction, can be further evaluated by biologists for experimental validation. In terms of future perspective, we are expecting that this method can be further improved and generalized in the first place to other plant-pathogenic bacteria and more.

Acknowledgements

A particular thanks goes to Paola Porracciolo for having developed MOnSTER and to all the supporting team.

References

- [1] Kube, M., Mitrovic, J., Duduk, B., Rabus, R. & Seemüller, E. Current view on phytoplasma genomes and encoded metabolism. *Sci. World J*, 2012.
- [2] Namba, S. Molecular and biological properties of phytoplasmas. *Proc. Japan Acad. Ser. B Phys. Biol. Sci.*, (95):401–418,2019.
- [3] Strauss, E. Phytoplasma research begins to bloom. *Science*, (325): 388–390, 2009.
- [4] Hoshi, A. et al. A unique virulence factor for proliferation and dwarfism in plants identified from a phytopathogenic bacterium. *Proc. Natl. Acad. Sci. U. S. A.*, (106): 6416–6421, 2009.
- [5] Huang, W. et al. Parasitic modulation of host development by ubiquitin-independent protein degradation. *Cell*, (184): 5201-5214.e12, 2021.
- [6] Pecher, P. et al. Phytoplasma SAP11 effector destabilization of TCP transcription factors differentially impact development and defence of Arabidopsis versus maize. *PLoS Pathog.*, (15): 1–27, 2019.
- [7] Bai, X. et al. AY-WB phytoplasma secretes a protein that targets plant cell nuclei. *Mol. Plant-Microbe Interact.*, (22): 18–30,2009.
- [8] MacLean, A. M. et al. Phytoplasma effector SAP54 induces indeterminate leaf-like flower development in Arabidopsis plants. *Plant Physiol.*, (157): 831–841, 2011.
- [9] Maejima, K. et al. Recognition of floral homeotic MADS domain transcription factors by a phytoplasmal effector, phyllogen, induces phyllody. *Plant J.*, (78): 541–554, 2014.
- [10] Liu, T. et al. Unconventionally secreted effectors of two filamentous pathogens target plant salicylate biosynthesis. *Nat. Commun.*, (5), 2014.
- [11] Garcion, C., Béven, L. & Foissac, X. Comparison of Current Methods for Signal Peptide Prediction in Phytoplasmas. *Front. Microbiol.*, (12): 1–17, 2021.
- [12] Kristianingsih, R. & MacLean, D. Accurate plant pathogen effector protein classification ab initio with deeppred: an ensemble of convolutional neural networks. *BMC Bioinformatics*, (22): 1–22, 2021.
- [13] Xu, C. & Jackson, S. A. Machine learning and complex biological data The revolution of biological techniques and demands for new data mining methods. *Genome Biol.*, (20): 1–4, 2019.
- [14] Vens, C., Rosso, M. N. & Danchin, E. G. J. Identifying discriminative classification-based motifs in biological sequences. *Bioinformatics*, (27): 1231–1238, 2011.
- [15] Bailey, T. L. STREME: accurate and versatile sequence motif discovery. *Bioinformatics*, (37): 2834–2840, 2021.
- [16] Asgari, E., McHardy, A. C. & Mofrad, M. R. K. Probabilistic variable-length segmentation of protein sequences for discriminative motif discovery (DiMotif) and sequence embedding (ProtVecX). *Sci. Rep.*, (9): 1–16, 2019.
- [17] Nielsen H., Predicting Secretory Proteins with SignalP. Protein Function Prediction. *Methods in Molecular Biology*, (1611): 59-73, 2017.
- [18] Krogh, A., Larsson, B., Von Heijne, G. & Sonnhammer, E. L. L. Predicting transmembrane protein topology with a hidden Markov model: Application to complete genomes. *J. Mol. Biol.*, (305): 567–580, 2001.
- [19] Necci, M., Piovesan, D., Dosztanyi, Z. & Tosatto, S. C. E. MobiDB-lite: Fast and highly specific consensus prediction of intrinsic disorder in proteins. *Bioinformatics*, (33): 1402–1404, 2017.
- [20] Sigrist, C. J. A. et al. New and continuing developments at PROSITE. *Nucleic Acids Res.*, (41): 344–347, 2013.
- [21] Seemüller, E., Sule, S., Kube, M., Jelkmann, W. & Schneider, B. The AAA+ ATPases and HflB/FtsH proteases of ‘Candidatus Phytoplasma mali’: Phylogenetic diversity, membrane topology, and relationship to strain virulence. *Mol. Plant-Microbe Interact.*, (26), 367–376, 2013.

Goodness of Fit for Bayesian Generative Models with Applications in Population Genetics

Guillaume LE MAILLOUX^{1,2}, Jean-Michel MARIN¹, Paul BASTIDE¹ and Arnaud ESTOUP²

¹ IMAG, University of Montpellier, Cnrs -- Montpellier, France

² CBGP, Univ Montpellier, CIRAD, INRAE, Institut Agro, IRD, Montpellier, France

Corresponding author: `guillaume.le-mailloux@umontpellier.fr`

Abstract *Population genetics often use complex, and hence intractable, generative models, that can be studied using likelihood-free simulation-based inferential methods such as Approximate Bayesian Computation (ABC). ABC starts by simulating a large number of datasets from a set of proposed models, each model encompassing an evolutionary scenario and a set of priors, and then compares the observed dataset to those simulated using adequate statistical measures, for e.g. model selection or parameter inference. Goodness-of-fit (GOF) methods aim at evaluating the level of adequacy between the observed dataset and a given model of interest, typically using an hypothesis-testing approach [1]. In an ABC context, this question can be re-framed as a novelty detection problem, in which one seeks to evaluate to which extent the observed dataset is an outlier compared to the simulated datasets. Many scores have been used as metrics to construct GOF test statistics and have been extensively tested in the literature [2]. Here we show that the approach proposed by [1] amounts to using a k -Nearest Neighbors (k -NN) metric, and we propose an alternative score based on the Local Outlier Factor (LOF) [3]. The LOF-based approach provides a substantially higher power than the k -NN-based one in controlled experiments based on toy as well as in complex population genetics models. We finally illustrate our LOF-based GOF method with a population genetics dataset composed of Single Nucleotide Polymorphism (SNP) markers to study a set of evolutionary scenarios of modern Human populations.*

Keywords generative models, goodness-of-fit, statistical tests, novelty detection, likelihood-free, population genomics, Single Nucleotide Polymorphism.

1 Introduction

In several disciplines such as population genetics, the amount of accessible data is growing, in size and in complexity, at an astonishing rate. Therefore, one has to develop complex models to try to explain all the subtleties of the data. This complexity comes at a cost, as the associated likelihood functions are often intractable. However, these models are often generative, and statistical methods relying on simulations, such as Approximate Bayesian Computation (ABC) [4], have been developed to circumvent this issue.

In a Bayesian setting, a model m is defined as a couple encompassing a prior distribution on the parameter space $\pi_m(\theta)$ and a likelihood function $f_m(z|\theta)$, so that $\mathbb{P}_m(z) = \int f_m(z|\theta)\pi_m(\theta)d\theta$. If the likelihood is intractable, one still can draw samples from it through simulations:

$$\theta_i \sim \pi_m(\theta), z_i \sim f_m(z|\theta_i), 1 \leq i \leq N.$$

For the vast majority of applications, summary statistics $\eta(z_i)$ are computed to reduce the dimensionality of the data. The simulated parameters and statistics are then gathered in a *reference table* (see Tab. 1), that can be used for further statistical analysis.

Goodness-of-fit (GOF) aims at testing whether an observed dataset z_{obs} is likely or not to have been generated by a given model m . Lemaire et al. [1] proposed an hypothesis testing procedure to test whether, for a given model, $\eta_{obs} = \eta(z_{obs})$ lies inside of the empirical distribution of the $\eta_i = \eta(z_i)$ from the reference table. It relies on the following test statistic:

$$D_m(\eta) = \frac{1}{\tilde{n}} \sum_{j=1}^{\tilde{n}} d(\eta_j, \eta), \quad (1)$$

$i = 1$	θ_1	z_1	$\eta(z_1)$
\dots	\dots	\dots	\dots
$i = N$	θ_N	z_N	$\eta(z_N)$

Tab. 1. A Reference table for a given Bayesian generative model m . Each row i is associated with a set of parameter θ_i drawn from a prior, generated data z_i and summary statistics $\eta_i = \eta(z_i)$.

where d is the euclidean distance and the set $\{\eta_j\}_{1 \leq j \leq \tilde{n}}$ is the set of the \tilde{n} nearest simulated summary statistics with η . In practice, they chose \tilde{n} as a given fraction of the total number of simulations N with the model m , for instance $\tilde{n}/n = 1\%$. In their words, “*finding an observed summary statistic outside of the range of the distribution is an indication of poor fit*”. Therefore, $D_m(\eta_{obs})$ is a goodness-of-fit statistic. Indeed, if this statistic is too large, it indicates that the explanatory power of the tested model m is not adapted to describe the observed data. More precisely, for a given model m , they report the p-value $\mathbb{P}_m(D_m(\eta(z)) \geq D_m(\eta(z_{obs})))$, with $z \sim \mathbb{P}_m$, that is conditional on the reference table through the definition of D_m . Unfortunately, this p-value has no closed form, but it can be estimated using simulations. Indeed, using H other summary statistics $\{\eta_h = \eta(z_h)\}_{1 \leq h \leq H}$, with $z_h \sim \mathbb{P}_m$, we have the usual empirical approximation:

$$\mathbb{P}_m(D_m(\eta(z)) \geq D_m(\eta(z_{obs}))) \approx \frac{1}{H} \sum_{h=1}^H \mathbb{1}(D_m(\eta_h) \geq D_m(\eta_{obs})). \quad (2)$$

If the result is less than, say, 5%, then we reject the model m for this observation z_{obs} . In this case, the test concludes that the model m of interest (i.e. the couple scenario - prior) needs to be improved.

2 A novelty detection algorithm: LOF

A crucial observation is that the GOF problem described above can be re-framed as a *novelty detection* problem. Indeed, the question is, given a dataset of points η_i from the reference table, that all come from distribution defined by the model m , can we detect whether η_{obs} is novel compared to them, i.e. whether it can be considered as an outlier in this dataset? This simple observation unlocks the use of the novelty detection literature, which has been steadily growing over the past few decades. Depending on the type of outliers considered, several reviews and benchmarks (see *e.g.*, [5,2]) are available to compare the many existing methods.

Having this in mind, the statistic defined in (1) can be recognized as the well-known mean k-Nearest Neighbors statistics (k-NN) with $k = \tilde{n}$, which is widely used for outlier detection (see *e.g.* [6]). In this paper, as an alternative to this k-NN statistics, we focus on the Local Outlier Factor (LOF, [3]) statistic. In preliminary tests (not shown here) the LOF indeed proved to be a reliable and easy to compute statistic, that only relies on a single and easy to understand hyperparameter (k). The original LOF algorithm computes a Local Outlier Factor for every point in an empirical distribution as shown in Fig. 1. It relies on the k-local reachable density in x with respect to the dataset $R = \{x_j\}$:

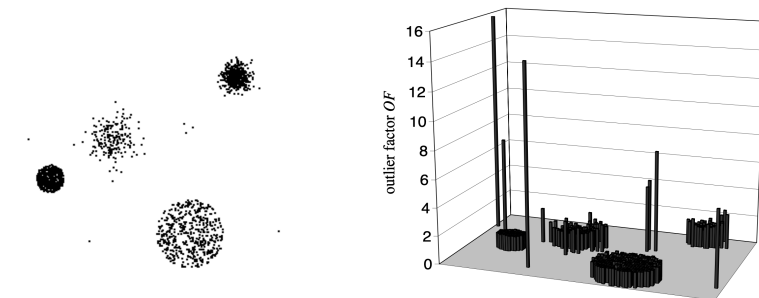


Fig. 1. Local Outlier Factor taken from [3]. Outliers need to be defined relatively to the local density of the points around them. It is worth noting that some clusters of points do not have the same local density, which could bias the statistic (1).

$$\text{lrd}_k(x; R) = \left(\frac{1}{k} \sum_{p \in N_k(x; R)} \text{reach-dist}(x, p) \right)^{-1}, \quad (3)$$

where $\text{reach-dist}(x, p) = \max(d(x, p), \text{k-dist}(p; R))$, with $\text{k-dist}(p; R)$ is the k^{th} nearest distance between p and points in R , and $N_k(x; R)$ are the k nearest points of x in R (excluding x).

The authors of LOF in [3] use the reach-dist rather than simply euclidean distance to stabilize the lrd. It is worth stressing that if we were to replace the reach-dist by the euclidean distance, then we would find an identity between (1) and (3) that is $\text{lrd}_k(x; R)^{-1} = D_m(x)$ with $k = \tilde{n}$ and R the set of summary statistics coming from model m in the reference table. This relationship means that the GOF statistic (1) is *local*. But, as illustrated in Fig. 1, the density of points in a distribution could sensibly vary. Therefore it does not seem appropriate to use directly D_m as a goodness-of-fit statistic. Indeed, some points can have a much smaller density than others without being outliers or novel points. In order to fix this issue of globally changing density, the LOF statistic is defined with a ratio as:

$$\text{LOF}_k(x; R) = \frac{\frac{1}{k} \sum_{p \in N_k(x; R)} \text{lrd}_k(p; R)}{\text{lrd}_k(x; R)}. \quad (4)$$

This statistic is fully local and can be nicely interpreted. For instance, $\text{LOF}_k(x; R) \approx 3$ means that the density in x is three times smaller than the average density of its neighborhood. Then, we can use the approach discussed in introduction to define a new goodness-of-fit measure that is the same as (2) with $D_m(\cdot)$ replaced by $\text{LOF}_k(\cdot; R)$:

$$\mathbb{P}_m \left(\text{LOF}_k(\eta(z); R) \geq \text{LOF}_k(\eta(z_{\text{obs}}); R) \right) \approx \frac{1}{H} \sum_{h=1}^H \mathbb{1} \left(\text{LOF}_k(\eta_h; R) \geq \text{LOF}_k(\eta_{\text{obs}}; R) \right).$$

Interestingly, the only hyperparameter of LOF is an integer k (often defaulting to ≈ 20 in practice), which defines the size of the neighborhood. In order to avoid the difficult choice of a "good" value for k , the authors in [3] suggest to define a Max-LOF statistic: $\text{Max-LOF}(x, R) = \max_{k \in I} \text{LOF}_k(x, R)$ where I is an interval of integers between, *e.g.*, 5 and 20.

3 Power Computations

In order to compare the goodness-of-fit measure by LOF and Mean k-NN, we compute a p-value as Eq. 2 for various models m with multiple pseudo-observed z_{obs} (≈ 1000 in practice) simulated from an other model \tilde{m} . We gather those different p-values to compute a power with a probability α of Type I error:

$$\text{Power} = \mathbb{P}_{z_{\text{obs}} \sim \tilde{m}} \left(\mathbb{P}_{z \sim m} \left[T(\eta(z), R) \geq T(\eta(z_{\text{obs}}), R) \right] \leq \alpha \right), \quad (5)$$

where $T(\cdot, R)$ is the goodness-of-fit statistic (either LOF or Mean k-NN). This statistical power is conditioned on the empirical distribution R from the reference table representing the distribution of model m .

Suppose $\tilde{m} \neq m$ and suppose also that this power is equal to 90%. This means that, when generating z_{obs} from \tilde{m} , the test correctly rejects the model m at the level α with a probability 0.9. Obviously (i.e. by construction), if $\tilde{m} = m$, the distribution of p-values is uniform, so we have $\text{Power} = \alpha$, and in this case the test wrongly rejected the model m with probability α . See Fig.2 Left panel for an illustration based on the toy example detailed in section 4.

To compute this power we use twice the same approximation as in Eq. 2. We note $\{\eta_h = \eta(z_h)\}_{1 \leq h \leq H}$ with $z_h \stackrel{iid}{\sim} \mathbb{P}_m$ and $\{\eta_{\text{obs}}^j = \eta(z_{\text{obs}}^j)\}_{1 \leq j \leq J}$ with $z_{\text{obs}}^j \stackrel{iid}{\sim} \mathbb{P}_{\tilde{m}}$, then this approximation follows:

$$\begin{aligned} \text{Power} &\approx \frac{1}{J} \sum_{j=1}^J \mathbb{1} \left(\frac{1}{H} \sum_{h=1}^H \mathbb{1} \left(T(\eta_h, R) \geq T(\eta_{\text{obs}}^j, R) \right) \leq \alpha \right) \\ &= \frac{1}{J} \#\{ j \mid 1 \leq j \leq J, T(\eta_{\text{obs}}^j, R) \geq T_{1-\alpha} \}, \end{aligned} \quad (6)$$

where $T_{1-\alpha}$ is the $1 - \alpha$ quantile of $\{T(\eta_h, R) | 1 \leq h \leq H\}$, and $\#$ indicates the cardinal of the set.

In practice, we use $H = J = 1000$ with a data set R of 10000 points. In the examples and illustrations detailed in sections 4 and 5, the probability α of a Type I error is set to 0.05.

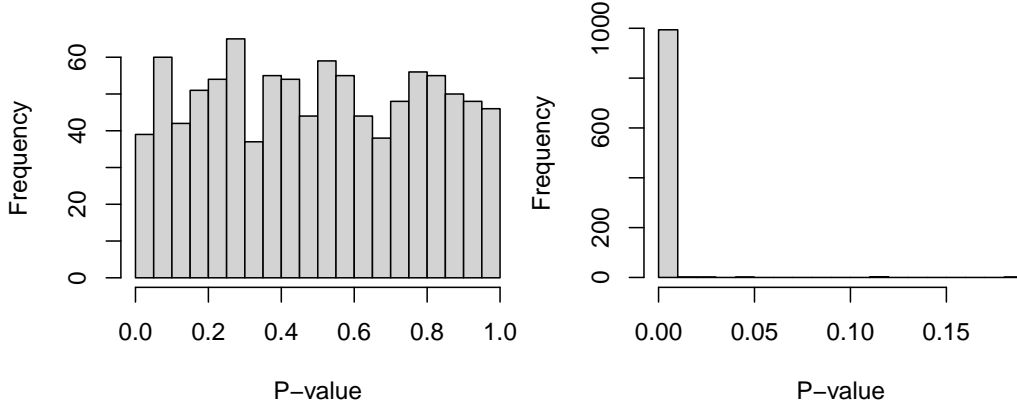


Fig. 2. Examples of distribution of p-values. *Left:* When $m = \tilde{m}$ the distribution is uniform. *Right:* Here $m = m_1$ Gaussian and $\tilde{m} = m_2$ Laplace, see section 4 for the definition of m_1 and m_2 . The more concentrated near zero the distribution is, the greater the power is.

4 Toy example

In this first example, we consider two Bayesian generative models that we will call Gaussian and Laplace respectively, defined as follows:

$$\begin{aligned} m_1 \text{ Gaussian: } \theta = (\mu, \sigma) &\sim \mathcal{U}(-5, 5) \otimes \mathcal{U}(1, 4), \text{ then } (z_i)_{1 \leq i \leq d} \stackrel{iid}{\sim} \mathcal{N}(\mu, \sigma^2) \\ m_2 \text{ Laplace: } \theta = (\mu, \sigma) &\sim \mathcal{U}(-5, 5) \otimes \mathcal{U}(1, 4), \text{ then } (z_i)_{1 \leq i \leq d} \stackrel{iid}{\sim} \mathcal{L}(\mu, \sigma^2) \end{aligned} \quad (7)$$

with densities $\mathbb{P}_{m_1}(z_i|\theta) = (2\pi\sigma^2)^{-\frac{1}{2}}e^{-\frac{1}{2}(z_i/\sigma)^2}$ and $\mathbb{P}_{m_2}(z_i|\theta) = (2\sigma^2)^{-\frac{1}{2}}e^{-\sqrt{2}|z_i/\sigma|}$, so that the mean and the standard deviations are μ and σ for both models.

The vector $z = (z_i)_{1 \leq i \leq d}$ represents our raw data that needs to be reduced by summary statistics, in practice $d = 350$. We choose 20 L-moments that are robust statistics of mean, standard deviation, kurtosis, skewness and higher order L-moments, computed thanks to the R library *lmom* [7].

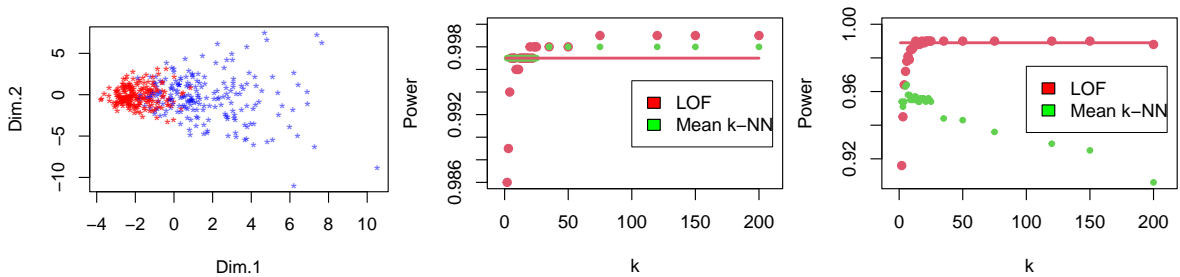


Fig. 3. *Left:* The two first axis of a PCA computed in the space of summary statistics, with red points generated from the Gaussian model and blue points from the Laplace model. *Center:* Power of the goodness-of-fit test for various k values. The pseudo-observed datasets comes from the Laplace model and the reference table is simulated with the Gaussian one. In this case, it is easy to conclude that the pseudo-observed datasets does not come from the tested model (cf. high power for both LOF and Mean k-NN). *Right:* The pseudo-observed datasets comes from the Gaussian model and the reference table is simulated with the Laplace one. In this case, it is harder to conclude that the pseudo-observed does not come from the tested model (cf. substantially lower power for Mean k-NN and to a lower extent for LOF). The red line is the power with the Max-LOF statistic.

In Fig. 3, one can see the first two axes of a PCA computed in the space of summary statistics for 1000 points from m_1 and 1000 points from m_2 . This shows that the support of the distribution

of the Gaussian model m_1 is included in the support of the distribution of the Laplace model m_2 . Therefore, the goodness-of-fit of the Gaussian model m_1 is expected to be often poor if a pseudo-observed dataset comes from the Laplace model m_2 , whereas if a pseudo-observed comes from the Gaussian model m_1 the goodness-of-fit of the Laplace model m_2 will often be good. This asymmetry explains why the power defined in (5) shown in Fig. 3 is greater when we compute the goodness-of-fit of datasets generated from the Laplace model with a reference table simulated according to the Gaussian model, than when we exchange those two models.

This first toy example shows that the LOF-statistic performs well. In the easy case (Center panel in Fig. 3), Mean k-NN is slightly better for some values of k , but in either case the power is greater than 0.99. In the more difficult case (Right panel in Fig. 3), LOF clearly outperforms Mean k-NN, and, excluding the very small values of k , the worst k -values for LOF give better results than the optimal k -value for Mean k-NN.

5 Example using population genetics models with simulated pseudo-observed SNP datasets

We considered a case study where one wants to oppose two evolutionary scenarios of four populations using 5000 single nucleotide polymorphism (SNP) genetic markers genotyped for 10 individuals per population. These two scenarios are depicted in Fig. 4. In scenario 1, populations diverged serially from each other, population 1 being the ancestral population (pop 1 \rightarrow pop 2 \rightarrow pop 3 \rightarrow pop 4). In scenario 2, populations 2, 3 and 4 diverged independently from the ancestral population 1 (pop 1 \rightarrow pop 2, pop 1 \rightarrow pop 3, and pop 1 \rightarrow pop 4).

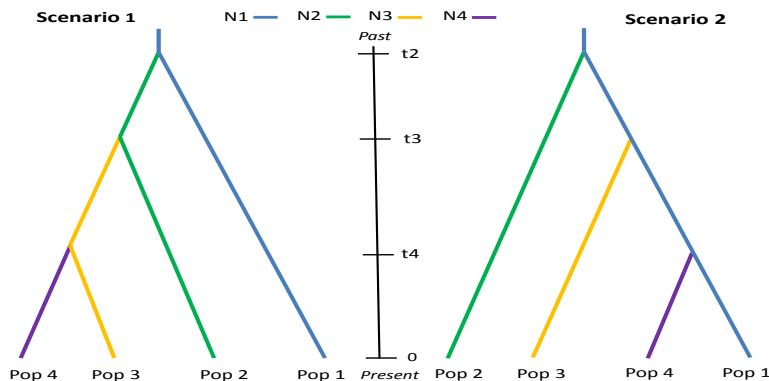


Fig. 4. Illustrations of the two evolutionary scenarios considered in the population genetics example of section 5. In Scenario 2, pop 1 is a common ancestral population.

These scenarios are often opposed when analyzing invasion or colonization histories, with scenario 1 corresponding to the case of a cascade of secondary invasion (colonization) events after a primary invasion (colonization) event and scenario 2 corresponding to the case of multiple independent invasion (colonization) events from a single source population. We used the package ABC Random Forest v1.0 [8] to simulate a reference table including 10000 simulated datasets for each scenario, using the following prior distributions for the historical and demographical parameters: $\mathcal{U}[1000; 10000]$ distributions for the effective population sizes N_1 , N_2 , N_3 and N_4 in number of diploid individuals, and $\mathcal{U}[10; 1000]$ for the divergence times t_2 , t_3 and t_4 in number of generations (with $t_2 > t_3$ and $t_3 > t_4$). For each scenario, we also simulated in the same manner a test dataset including 1000 pseudo-observed datasets. The observed and simulated SNP datasets were summarized using a total of 130 summary statistics describing genetic variation within populations (*e.g.* proportion of monomorphic loci, heterozygosity, population-specific F_{ST}) and between pair, triplet or quadruplet of populations (*e.g.*, Nei's distance, F_{ST} -related statistics, Patterson's allele-sharing f -statistics, coefficients of admixture) to describe genetic variation among various population combinations (see [8] for details). A simple PCA analysis processed on the summary statistics values generated under the scenario 1 and scenario 2 indicates that the two datasets are only mildly overlapping (Fig. 5 Left panel).

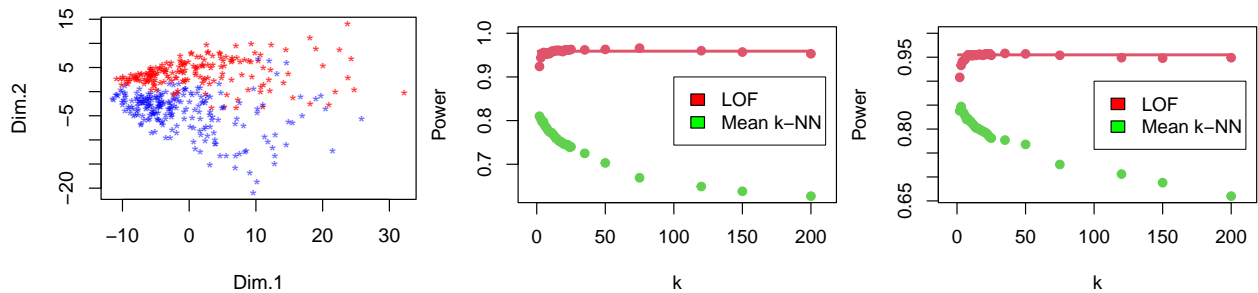


Fig. 5. *Left:* First two axes of a PCA computed in the space of summary statistics. 1000 blue points come from scenario 1 and 1000 red points of summary statistics come from scenario 2. *Center and Right panels:* Power of the goodness-of-fit test for various k values. *Center panel:* The pseudo-observed comes from the scenario 2 and the reference table is simulated with the scenario 1. *Right panel:* The pseudo-observed comes from the scenario 1 and the reference table is simulated with the scenario 2. The red line is the power with the Max-LOF statistic.

This second example nicely illustrates the superiority of LOF over Mean k -NN statistic for goodness-of-fit purposes in a population genetics context. The statistical power is always significantly higher with LOF than with Mean k -NN. Furthermore, the value of k does not seem crucial as indicated by [3] and the power of Max-LOF is nearly optimal.

6 Illustration on a real SNP dataset of Human populations

We applied our goodness-of-fit methodology on a real dataset including 24,690 independent SNP markers (with a minimum allele frequency of 0.01) genotyped in four Human populations (30 individuals per populations) by The 1000 Genomes Project Consortium (2012) [9]. The four populations include Yoruba (Africa), Han (East Asia), British (Europe) and American individuals of African ancestry. Six scenarios of evolution of the four Human populations were compared (Fig. 6). Following [8], we used the package ABC Random Forest v1.0 to simulate 10000 points per scenario. The observed and simulated SNP datasets were summarized using 130 summary statistics. See [8] for details regarding SNP data and summary statistics features. See also Fig. 6 legend for the prior distributions of demographical parameters used for simulation. A simple PCA analysis processed on the summary statistics values generated under the six different scenario-prior couples indicate that the simulated datasets are substantially overlapping, including in the area around the observed dataset (not shown here).

Scenario	1	2	3	4	5	6
GOF by Max-LOF	0 ± 0.001	0.088 ± 0.009	0.004 ± 0.002	0 ± 0.001	0.002 ± 0.001	0 ± 0.001
GOF by Mean k -NN	0 ± 0.001	0.258 ± 0.013	0.077 ± 0.008	0 ± 0.001	0.018 ± 0.004	0.018 ± 0.004

Tab. 2. P-values for each of the six proposed scenarios of Human evolution estimated for a real SNP dataset. We used $H = 1000$ and standard errors on p-values were computed as $\sqrt{p(1-p)/H}$. We use $k=1$ as previous examples suggested that it is the more powerful value for GOF by Mean k -NN.

A Random Forest algorithm, processed as detailed in [8], selected the scenario 2 as the best forecast scenario with a posterior probability of 0.987. Considering previous population genetics studies in the field, it is not surprising that scenario 2, which includes a single out-of-Africa colonization event giving an ancestral out-of-Africa population with a secondary split into one European and one East Asian population lineage and a recent genetic admixture of Americans of African origin with their African ancestors and European individuals, was selected (*e.g* [10]). Using the observed dataset as target, we then applied our GOF methodologies on the six scenario-prior couple using 10 000 simulated datasets for each couple. Results are summarized in Table 2. GOF values are greater than 5% , i.e. 25.8% for GOF by Mean k -NN and 8.8% for GOF by Max-LOF, indicating an absence of major misfit between

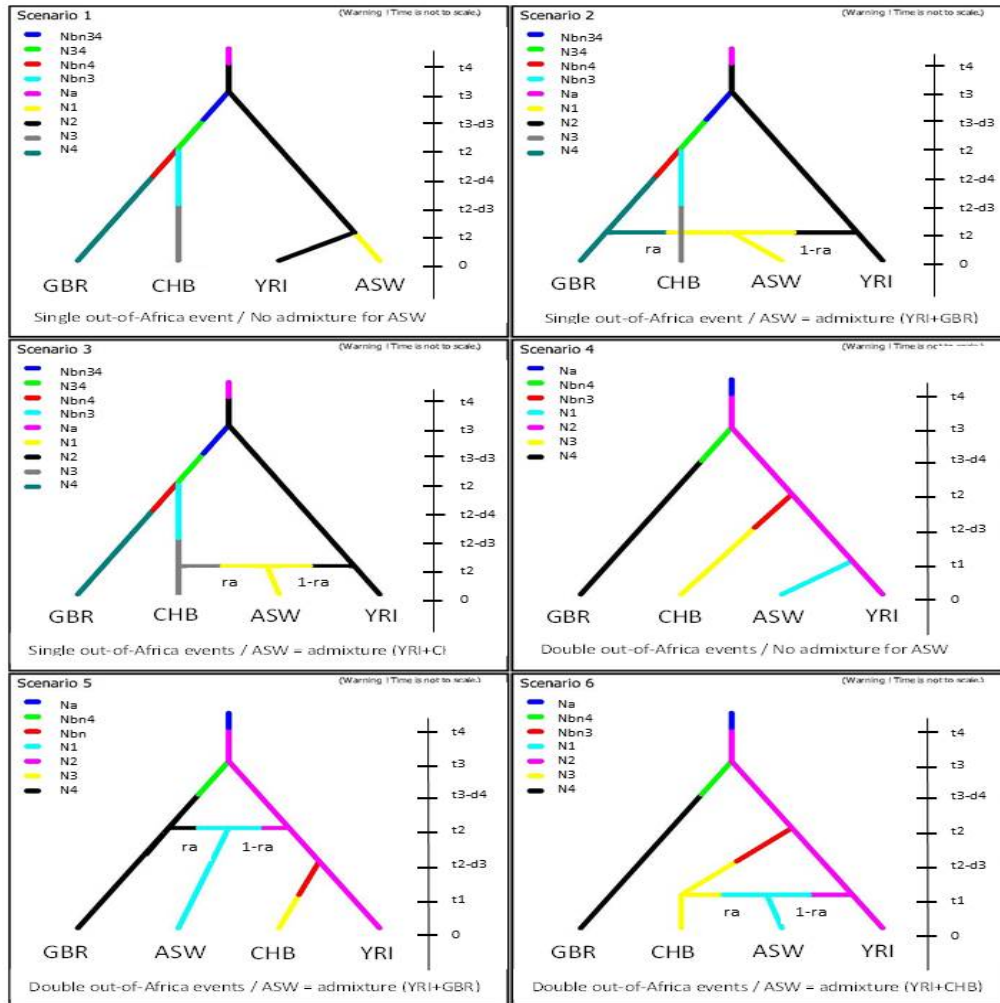


Fig. 6. Six scenarios of evolution of four modern Human populations. The genotyped populations are YRI = Yoruba (Nigeria, Africa), CHB = Han (China, East Asia), GBR = British (England and Scotland, Europe), and ASW = Americans of African Ancestry (USA). The six scenarios differ from each other by one ancient and one recent historical event: (i) a single out-of-Africa colonization event giving an ancestral out-of-Africa population which secondarily split into one European and one East Asian population lineage (scenarios 1, 2 and 3), versus two independent out-of-Africa colonization events, one giving the European lineage and the other one giving the East Asian lineage (scenarios 4, 5 and 6). (ii) The possibility (or not; scenarios 1 and 4) of a recent genetic admixture of ASW individuals with their African ancestors and individuals of European (scenarios 2 and 5) or East Asia (scenarios 3 and 6) origins. The prior distributions of the parameters used to simulate SNP datasets are as followed: $\mathcal{U}[100; 10000]$ for the split times t_2 and t_3 (in number of generations), $\mathcal{U}[1; 30]$ for the admixture (or split) time t_1 , $\mathcal{U}[0.05; 0.95]$ for the admixture rate ra (proportion of genes with a non-African origin; only for scenarios with admixture), $\mathcal{U}[1000; 100000]$ for the stable effective population sizes N_1 , N_2 , N_4 , N_4 and N_{34} (in number of diploid individuals), $\mathcal{U}[5; 500]$ Uniform[5; 500] for the bottleneck effective population sizes N_{bn3} , N_{bn4} , and N_{bn34} , $\mathcal{U}[5; 500]$ for the bottleneck durations d_3 , d_4 , and d_{34} , $\mathcal{U}[100; 10000]$ for both the ancestral effective population size N_a and the time of change to N_a . Conditions on time events were $t_4 > t_3 > t_2$ for scenarios 1, 2 and 3, and $t_4 > t_3$ and $t_4 > t_2$ for scenarios 4, 5 and 6.

the observed dataset and the selected scenario 2-prior couple. In agreement with the higher power of the GOF by LOF method revealed in our previous simulation-based studies, GOF values were globally lower when using the GOF by LOF method. GOF p-values were particularly low (i.e. $< 5\%$) for five scenarios (i.e. scenarios 1, 3, 4, 5 and 6) using the GOF by LOF method and for four scenarios (i.e.

scenarios 1, 4, 5, 6) using the GOF by Mean k-NN method, revealing the presence of major misfits between the observed dataset and those simulated with these scenario-prior couples.

7 Discussion

The hypothesis-testing GOF approach proposed by [1] could theoretically be adapted to any outlier or novelty detection algorithm. Here, we studied the LOF algorithm because of its simplicity and the absence of fine tuned hyperparameters. In agreement with this, we observed that, for all tests we performed, the optimal k value were always around 15 and that the power is globally not dramatically influenced by the precise value of k. We found that LOF provides better result than Mean k-NN. This is because the density of a complicated distribution may change in the big summary statistic space. This suggests that it might be of interest to develop and test (even) more complex outlier scores.

It is worth (re)stressing here that we tested the adequacy between observed summary statistics and a *marginal* distribution \mathbb{P}_m that depends on a prior distribution for parameters. Consequently, this GOF methodology depends on the prior distributions of the scenario - prior couple, a feature confirmed by several test we processed using different prior sets (results not shown). Practitioners might (also) want to validate their selected models *and* the inferred (posterior) parameters. We are currently working on the extension of our GOF methodologies to test a couple scenario - posterior instead of a scenario - prior couple.

Finally, simulations of datasets from complex models such as those used in population genetics are often computationally expensive. One would hence save a lot of computation time and resources if one could early identify models that are really off relatively to the observed dataset of interest from a low number of simulations. As a matter of fact, if the GOF p-value computed from a low number of simulations turned out to be low for some of the compared model, then it would not be worth simulating additional datasets for those models for the following inferential steps (model choice and parameter estimation). The GOF by LOF methodology we propose also provide an efficient tool for this purpose.

References

- [1] Louisiane Lemaire, Flora Jay, I-Hung Lee, Katalin Csilléry, and Michael G. B. Blum. Goodness-of-fit statistics for approximate bayesian computation. January 2016.
- [2] Songqiao Han, Xiyang Hu, Hailiang Huang, Mingqi Jiang, and Yue Zhao. Adbench: Anomaly detection benchmark. In *Neural Information Processing Systems (NeurIPS)*, 2022.
- [3] Markus M. Breunig, Hans-Peter Kriegel, Raymond T. Ng, and Jörg Sander. LOF. *ACM SIGMOD Record*, 29(2):93–104, may 2000.
- [4] Jean-Michel Marin, Pierre Pudlo, Christian P. Robert, and Robin J. Ryder. Approximate bayesian computational methods. *Statistics and Computing*, 22(6):1167–1180, October 2012.
- [5] Markus Goldstein and Seiichi Uchida. A Comparative Evaluation of Unsupervised Anomaly Detection Algorithms for Multivariate Data. *PLOS ONE*, 11(4):e0152173, April 2016.
- [6] Sridhar Ramaswamy, Rajeev Rastogi, and Kyuseok Shim. Efficient algorithms for mining outliers from large data sets. *ACM SIGMOD Record*, 29(2):427–438, May 2000.
- [7] J. R. M. Hosking. *L-Moments*, 2022. R package, version 2.9.
- [8] François-David Collin, Ghislain Durif, Louis Raynal, Eric Lombaert, Mathieu Gautier, Renaud Vitalis, Jean-Michel Marin, and Arnaud Estoup. Extending approximate bayesian computation with supervised machine learning to infer demographic history from genetic polymorphisms using DIYABC random forest. *Molecular Ecology Resources*, 21(8):2598–2613, may 2021.
- [9] The 1000 Genomes Project Consortium. An integrated map of genetic variation from 1,092 human genomes. *Nature*, 491(7422):56–65, November 2012.
- [10] Katarzyna Bryc, Eric Y. Durand, J. Michael Macpherson, David Reich, and Joanna L. Mountain. The genetic ancestry of african americans, latinos, and european americans across the united states. *The American Journal of Human Genetics*, 96(1):37–53, jan 2015.

Acknowledgements

We thank Michael Blum for useful discussions about the GOF by Mean k-NN method. This work was supported by a fund from the French Agence National pour la Recherche (ANR projects GANDHI).

Session 4: Functional and integrative genomics

Alteration of ribosome function upon 5-fluorouracil treatment favors cancer cell drug-tolerance

Gabriel THERIZOLS^{1,5,7}, Zeina BASH-IMAM^{1,5,7}, Baptiste PANTHU², Christelle MACHON^{1,5,7}, Anne VINCENT^{1,5,7}, Julie RIPOLL³, Sophie NAIT-SLIMANE^{1,5,7}, Mounira CHALABI-DCHAR^{1,5,7}, Angéline GAUCHEROT^{1,5,7}, Maxime GARCIA^{1,5,7}, Florian LAFORETS^{1,5,7}, Virginie MARCEL^{1,5,7}, Jihane BOUBAKER-VITRE⁴, Marie-Ambre MONET^{1,5,7}, Céline BOUCLIER⁴, Christophe VANBELLE^{1,5,7}, Guillaume SOUAHLIA^{1,5,7}, Elise BERTHEL^{1,5,7}, Marie Alexandra ALBARET^{1,5,7}, Hichem MERTANI^{1,5,7}, Michel PRUDHOMME⁵, Martin BERTRAND⁵, Alexandre DAVID⁶, Jean-Christophe SAURIN^{1,5,7}, Philippe BOUVET^{1,5,7}, Eric RIVALS³, Théophile OHLMANN², Jérôme GUITTON^{1,4}, Nicole DALLA VENEZIA^{1,5,7}, Julie PANNEQUIN⁶, Frédéric CATEZ^{1,5,7} and Jean-Jacques DIAZ^{1,5,7}

¹ Inserm U1052, CNRS UMR5286 Centre de Recherche en Cancérologie de Lyon, F-69000, Lyon, France

² CIRI-Inserm U1111, Ecole Normale Supérieure de Lyon, Lyon, F-693643, France

³ LIRMM, Université Montpellier, F-34095, Montpellier, France

⁴ Laboratoire de toxicologie, Faculté de pharmacie de Lyon, Université de Lyon, 8 avenue Rockefeller, F-69373 Lyon, France

⁵ Centre Léon Bérard, F-69008 Lyon, France.

⁶ IGF, Univ. Montpellier, CNRS, INSERM, Montpellier, France.

⁷ Univ. Lyon 1, F-69000 Lyon, France.

Corresponding author: eric.rivals@lirmm.fr

Reference paper: Therizols *et al.* (2022) Alteration of ribosome function upon 5-fluorouracil treatment favors cancer cell drug-tolerance. *Nature Communications*. <http://dx.doi.org/10.1038/s41467-021-27847-8>

Mechanisms of drug-tolerance remain poorly understood and have been linked to genomic but also to non-genomic processes. 5-fluorouracil (5-FU), the most widely used chemotherapy in oncology is associated with resistance. While prescribed as an inhibitor of DNA replication, 5-FU alters all RNA pathways.

Here, we show that 5-FU treatment leads to the production of fluorinated ribosomes exhibiting altered translational activities. 5-FU is incorporated into ribosomal RNAs of mature ribosomes in cancer cell lines, colorectal xenografts, and human tumors. Fluorinated ribosomes appear to be functional, yet, they display a selective translational activity towards mRNAs depending on the nature of their 5'-untranslated region. The translational selectivity was determined by bioinformatic analysis of deep sequencing data that combine RNA-seq of both the cytosomal and polysomal fractions of mRNAs.

As a result, we find that sustained translation of IGF-1R mRNA, which encodes one of the most potent cell survival effectors, promotes the survival of 5-FU-treated colorectal cancer cells. Altogether, our results demonstrate that *man-made* fluorinated ribosomes favor the drug-tolerant cellular phenotype by promoting translation of survival genes.

This work has been published in 2022 in NATURE COMMUNICATIONS [1].

Acknowledgements

The project was supported by the *Ligue Contre le Cancer*, by the Institut National du Cancer (INCa, grant FluoRib - PLBIO18-131).

References

- [1] Gabriel Therizols, Zeina Bash-Imam, Baptiste Panthu, Christelle Machon, Anne Vincent, Julie Ripoll, Sophie Nait-Slimane, Mounira Chalabi-Dchar, Angéline Gaucherot, Maxime Garcia, Florian Laforêts, Virginie Marcel, Jihane Boubaker-Vitre, Marie-Ambre Monet, Céline Bouclier, Christophe Vanbelle, Guillaume Souahlia, Elise Berthel, Marie Alexandra Albaret, Hichem Mertani, Michel Prudhomme, Martin Bertrand, Alexandre David, Jean-Christophe Saurin, Philippe Bouvet, Eric Rivals, Théophile Ohlmann, Jérôme Guitton, Nicole Dalla Venezia, Julie Pannequin, Frédéric Catez, and Jean-Jacques Diaz. Alteration of ribosome function upon 5-fluorouracil treatment favors cancer cell drug-tolerance. *Nature Communications*, 13(1):173, January 2022.

Comparative analysis of whole blood transcriptomics between European and local Caribbean pigs in response to feed restriction in a tropical climate

Nausicaa Pouillet¹, Alice Choury¹, Oriane Devarieux¹, David Beramice², Laurent Dantec², Yoann Félicité¹,

Dalila Feuillet¹, Jean-Luc Gourdine¹ & Jean-Christophe Bambou¹

¹ASSET, INRAE, 97170 Petit-Bourg (Guadeloupe), France

²PTEA, INRAE, 97170 Petit-Bourg (Guadeloupe), France

Corresponding Author: nausicaa.pouillet@inrae.fr

Abstract *Feed restriction occurs frequently during pig growth, either due to economic reasons or stressful environmental conditions. Local breeds are suggested to have better tolerance to periods of feed restriction. However, the mechanisms underlying the response to feed restriction in different breeds is largely unknown. The aims of the present study were 1) to compare the transcriptome profile in response to feed restriction and refeeding of two contrasted breeds, Large White (LW), which has been selected for high performance, and Creole (CR), which is adapted to tropical conditions, and 2) to investigate the effect of a moderate feed restriction and refeeding on whole blood transcriptome. Analysis of blood transcriptome allows to study the response to feed restriction and refeeding in a dynamic way. RNAseq was performed on blood samples of growing LW and CR pigs at two time points: after 3 weeks of feed restriction and after 3 weeks of refeeding. The data was compared with samples from control animals offered the same diet on an ad libitum basis throughout the whole experiment. In terms of performance, CR pigs were less impacted by feed restriction than LW. The transcriptional response to feed restriction and refeeding between CR and LW was contrasted both in terms of number of DEGs and enriched pathways. CR demonstrated a stronger transcriptional response to feed restriction whereas LW had a stronger response to refeeding. Differences in the transcriptional response to feed restriction between CR and LW were related to cell stress response (Aldosterone Signalling, Protein ubiquitination, Unfolded Protein Signalling) whereas after refeeding, differences were linked to thermogenesis, metabolic pathways and cell proliferation (p38 MAPK, ERK/MAPK pathway). In both breeds, transcriptional changes related to the immune response were found after restriction and refeeding. Altogether, the present study indicates that blood transcriptomics can be a useful tool to study differential genetic response to feed restriction in a dynamic way. The results indicate a differential response of blood gene expression to feed restriction and refeeding between breeds, affecting biological pathways that are in accordance with performance and thermoregulatory results.*

Keywords Blood transcriptome, feed restriction, refeeding, Creole pig, tropical climate

1. Introduction

During the growing period, pigs may encounter periods of feed restriction due to economic reasons or environmental factors. When facing stressful environmental conditions, such as heat stress, poor sanitary conditions, social stress or disease pressure, pigs reduce their feed intake, leading to feed restriction [1]–[3]. During these periods of feed restriction, the growing pig must adjust its metabolism to maintain homeostasis through changes in nutrient partitioning between growth and maintenance. The animal responses to feed restriction is highly variable within and between populations and part of this variability may have a genetic basis [4], [5]. Our previous work compared the effect of feed restriction on two contrasted breeds, the Creole (CR) breed, a local breed well adapted to tropical conditions and that has not been submitted to genetic selection, and the Large White breed (LW) that has been selected for high growth performance in optimal conditions [6]. Our results suggested that the CR breed may be more tolerant to feed restriction.

In the context of climate change, there is a crucial need of information on local breeds and on their adaptation to specific environmental conditions, as they constitute genetic resources that are essential to maintain livestock systems diversity and ensure food security [7]. The CR breed provides a good model to study the genetic variability in the response to feed restriction in pigs [6], [8], [9].

Advances in high-throughput technologies such as transcriptomics offer opportunities to better understand complex biological mechanisms and to better characterize local breeds lacking this kind of data. The collection of blood samples is relatively easy compared to other tissues and provides the possibility of sampling the same animal at different time points. It is also a technique that would be easily transferable in breeding schemes. A recent study on divergent selected lines of pigs showed that the blood transcriptome is relevant to identify biological processes affected by genetic selection and feeding strategies [10]. In the present study, we used whole blood transcriptome analysis to better understand the molecular mechanisms underlying the differential breed response to feed restriction. The objectives of the current study were 1) to investigate the effect of a moderate feed restriction and refeeding on whole blood transcriptome, 2) to compare the transcriptome profile of two contrasted breeds, CR and LW, in response to feed restriction and refeeding.

2. Methods

All measurements and observations on animals were performed in accordance with the current law on animal experimentation and ethics. The French Ministry of Agriculture authorized the experiment referenced at n°APAFIS#18576-2019011614325318 (after the revision of the Animal Care and Use Committee of French West Indies and French Guyana) on living animals at the INRAE facility under the direction of N. Minatchy (INRAE-PTEA).

2.1. Animals and experiment design

A total of 30 growing pigs (15 LW and 15 CR) of the same age, with an average BW of 32.3 ± 1.7 kg for LW and 18.2 ± 1.0 kg for CR, were used for the experiment in the semi-open front building of the INRAE experimental farm located in Guadeloupe, French West Indies. At 12 weeks of age, pigs were allotted to 2 or 3 pens with a density of 10 pigs/pen (5 LW and 5 CR).

The experiment consisted of three consecutive periods. Period 1 (**P1**) was the initial period (7 days) where all pigs were fed *ad-libitum*. Period 2 (**P2**) was a 3-week period during which feed restriction was imposed to specific pens. Due to experimental limitations, the two feeding treatments were not balanced in number of animals. During P2, one pen (referred to as NF, 5 LW and 5 CR) continued to be fed *ad libitum*, whereas 2 pens (referred to as RF, 10 LW and 10 CR) had restricted access to the automatic feeder (from 7:00 to 17:00). Period 3 (**P3**) constituted the following 3-week period and corresponded to the refeeding period during which all animals were fed *ad libitum*.

2.2. Measurements

Blood samples were collected at the end of P2 (week 15) and at the end of P3 (week 21) at 08:00 in the morning. Jugular vein blood was obtained (10-mL BD K₂ EDTA Vacutainers tubes (BD, Franklin Lakes, NJ)) via venepuncture. For samples dedicated to RNA extraction, one volume of blood sample was mixed with one volume of lysis buffer from the Nucleospin RNA blood kit (Macherey-Nagel, Lyon, France). The obtained mixture was then stored at -80°C for later analyses.

2.3. RNA extraction and quality analysis

Total RNA was extracted from frozen blood samples of 28 animals from the first replicate [9 NF (4 CR, 5LW) and 19 RF (10 LW, 9 CR)] using the NucleoSpin RNA isolation kit (Macherey-Nagel, Hoerd, France) in accordance with the manufacturer's instructions. The total RNA concentration was measured with NanoDrop 2000 (ThermoScientific TM, France) and the quality was quantified using an Agilent 2100 Bioanalyzer (Agilent Technologies, France). The extracted total RNA was stored at -80°C until use.

2.4. Library preparation and sequencing

High-quality RNA (RIN > 7.5) was used for the preparation of cDNA libraries according to Illumina's protocols (Illumina TruSeq RNA sample prep kit for mRNA analysis). Briefly, poly-A mRNA was purified from 4µg of total RNA, fragmented and randomly primed for reverse transcription to generate double stranded cDNA. The cDNA fragments were then subjected to an end repair process, consisting of the addition of a single 'A' base, and the ligation of indexed Illumina adapters at both ends of cDNA. These products were then purified and enriched by PCR to create the final bar-coded cDNA library. After quality control and quantification, cDNA libraries were sequenced on 2 lanes on the NovaSeq6000 S4 (Illumina® NEB, USA) to obtain approximately 48 million reads (100 bp paired-end) for each sample.

2.5. Quality control and read mapping to the reference genome

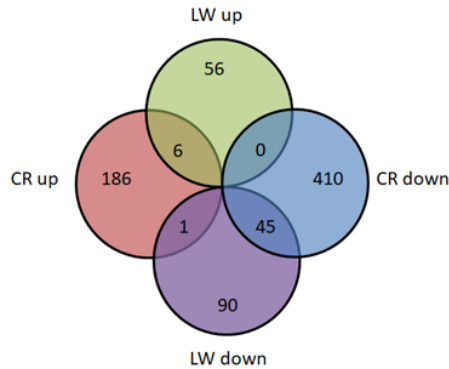
The quality control check on raw reads in FASTQ format were processed using FASTQC and the Q30, GC content and length distribution of the clean data were calculated. The sequences obtained by RNA-Seq were splice-aligned for each library, using STAR (version 2.3.0e with standard parameters) [11]. The reads were mapped to the *Sus Scrofa* genome (assembly 11.1). HTSeq (<http://pypi.python.org/pypi/HTSeq>) [12] was used to calculate the number of sequence reads aligned to all protein-coding genes from the ENSEMBL v74 annotation of the *Sus scrofa* genome. The Bioconductor package DeSeq2 [13] was then used to identify differentially expressed genes (DEGs). Two treatment comparisons were tested for DEGs for each breed: (i) RF v. NF at the end of Period 2; (ii) RF v. NF at the end of Period 3. Statistically significant ($P < 0.05$) DEGs with a Benjamini-Hochberg false discovery rate of < 0.05 were deemed to be significant. Analysis of canonical pathways and regulatory effects as well as network analysis were performed using Ingenuity pathway analysis (IPA) software (Ingenuity Systems, Redwood City, CA) for DEGs in each comparison. IPA identifies known regulators, including genes and other molecules that may affect the expression of DE genes, then it calculates a z-score, which is a statistical measure of the match between the expected relationship direction between the regulator and its targets, and the observed gene expression [14]. Moreover, KEGG pathway and Gene Ontology enrichment analyses were performed using ShinyGO [15].

3. Results

3.1. mRNA read alignment and differential gene expression

Following the period of feed restriction, at the end of P2, 648 genes were differentially expressed (DE) in CR, whereas 198 were DE in LW (Figure 1a). Of the 648 DEG in CR, 193 were up-regulated and 455 down-regulated. In LW, of the 198 DEG, 62 up-regulated and 136 down-regulated. CR and LW shared 51 DEGs in response to feed restriction, with 45 down-regulated and 6 up-regulated. Following refeeding, the opposite pattern was found, with a higher number of DEG in LW than CR (1538 in LW vs. 187 in CR) (Figure 1b). After refeeding, in both breeds, the majority of DEG were up-regulated (55% and 61% upregulated, in LW and CR, respectively) whereas after restriction, DEG were mostly down-regulated (69% and 70% downregulated, in LW and CR, respectively). Few DEG were shared by both breeds, with 28 upregulated and 19 down-regulated. An additional 21 genes were shared by both breeds but the direction of the fold change was reversed between the two breeds.

a. RF vs. NF, after restriction (P2)



b. RF vs. NF, after refeeding (P3)

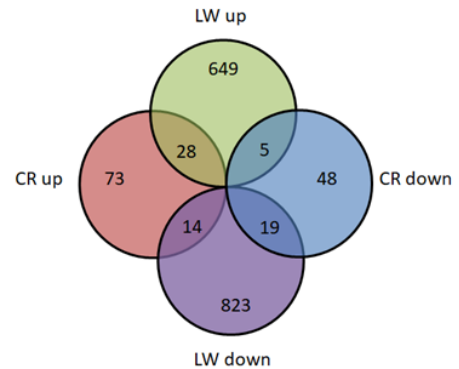
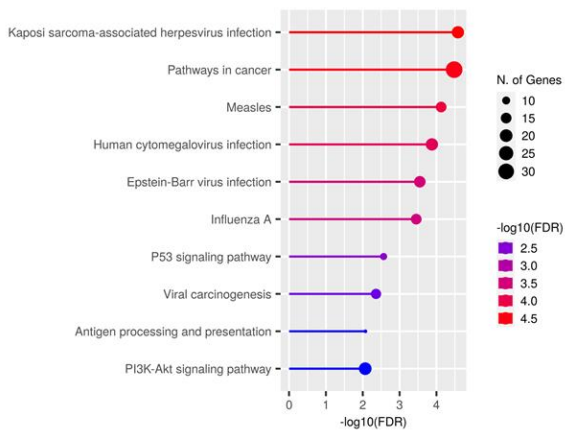


Figure 1. Venn diagrams displaying the number of differentially expressed genes (DEG) in Large White (LW) and Creole (CR) pigs for each comparison. RF: Restricted Feeding, NF: Normal Feeding. P2: restriction period, P3: refeeding period. Numbers in overlapping areas represent DEGs shared by both breeds.

3.2. Gene Ontology and Pathway analysis

The DEG from each comparison were submitted to ShinyGO [15] for Gene Ontology (GO) analysis. Pathway analysis based on the KEGG database revealed 39 enriched pathways at the end of P2 for CR and 18 at the end of P3 for LW (Top 10 shown in Fig. 2). However, the smaller number of DEG identified at the end of P2 for LW and at the end of P3 for CR did not allow to reach any significant KEGG pathway enrichment.

CR, RF vs. NF, after restriction (P2)



LW, RF vs. NF, after refeeding (P3)

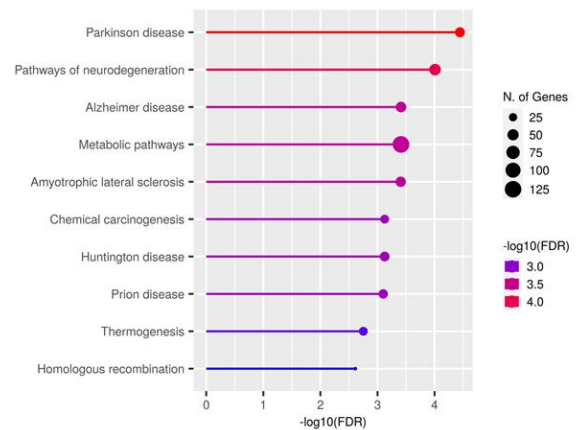


Figure 2. Top 10 significant KEGG pathways identified by ShinyGO [15]–[17] using DE genes between treatments. CR: Creole, LW: Large White, RF: Restricted Feeding, NF: Normal Feeding.

3.3. Ingenuity Pathway Analysis (IPA)

After feed restriction, at the end of P2, Ingenuity Pathway Analysis (IPA) identified 29 significant canonical pathways for LW and 179 for CR. Whereas after refeeding, at the end of P3, IPA found 30 canonical pathways for LW and 27 for CR. IPA was also used to compare results from the different comparisons in the 2 breeds between treatments (NF vs. RF), over time (after restriction and after refeeding). The top 10 canonical pathways and the top 10 diseases and biological functions were compared (Figure 3). When comparing the 2 breeds after restriction, synaptogenesis signalling was the only pathway to be significantly inhibited (z-score < 2) in both breeds and it was no longer inhibited after refeeding. In CR, after restriction, enriched pathways were inhibited and mostly related to the immune response (natural cell killer signalling, neuroinflammation

signalling, production of nitric oxide). When comparing results after restriction and after refeeding, all pathways and disease and biological functions had a z-score closer to 0 (lower activation) after P3 than after P2. For disease and biological functions, “organismal death”, “anemia”, “polycythemia” were activated in both breeds after restriction but it was no longer the case after refeeding. “Quantity of lymphocytes” was inhibited in both breeds after restriction. After refeeding, “quantity of lymphocytes” was still inhibited in LW to a lower extent but not in CR. “Immune response of cells” was inhibited in CR after restriction and to a lower extent after refeeding.

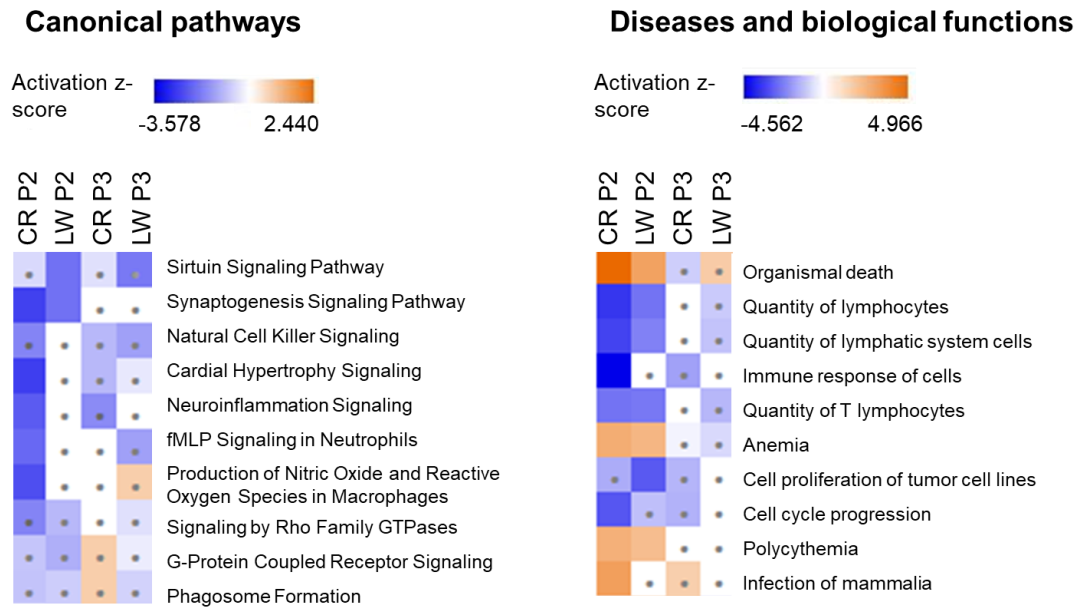


Figure 3. Heat map of canonical pathways and diseases and biological functions identified by Ingenuity Pathway Analysis using DE genes between treatments (RF vs. NF). CR: Creole, LW: Large White, RF: Restricted Feeding, NF: Normal Feeding. P2: restriction period, P3: refeeding period. Squares with dots indicates pathways for which activation/inhibition was not significant (z-score <|2|).

4. Discussion

Periods of feed restriction may occur during pig growth due to economic reasons or external factors, such as heat waves, inflammatory stress, feed transition or social stress [3]. Few studies have investigated the effect of feed restriction and refeeding on livestock transcriptome [18]–[20] and to our knowledge, there is no comparative analysis of the transcriptomic response to feed restriction and refeeding in different pig breeds. The present study aimed to investigate the effect of feed restriction and refeeding on the blood transcriptome of growing pigs from two contrasted breeds.

RNAseq analysis comparing the two feeding groups (RF vs. NF) show that after restriction there were more DEGs in CR than LW, suggesting that the response elicited by feed restriction is stronger in CR than LW. Consequently, after restriction, we also identified more enriched pathways in GO and IPA analysis for CR than LW. KEGG enrichment showed that the main pathways triggered after feed restriction in CR were related to immunity. Similar results were found after IPA analysis regarding canonical pathways after restriction in CR. The most enriched pathways were related to the immune response and viral infection (Interferon signalling, Th1 pathway), cancer (Pancreas adenocarcinoma signalling, Rac signalling) and Ephrin receptor signalling, which is involved in the maintenance of several processes including angiogenesis, stem cell differentiation and cancer. Finding many genes related to immunity in the blood transcriptome is not surprising as blood cells constitute one of the first lines of immune defence [21]. Similar findings were found in pig studies on blood transcriptome response to genetic selection for feed efficiency and nutritional status [10], [22]. Moreover, genes involved in the immune response were also found to be differentially expressed after dietary restriction in beef cattle jejunal epithelium [18]. Reports in mice, human and rats have also described improved immune

function after periods of caloric restriction [23]–[25]. The main hypothesis is that the immune response may be involved in nutrient partitioning, allowing activation of tissue mobilisation during dietary restriction [26]. GO analysis for LW after restriction comparing RF to NF did not allow to reach any enrichment, probably due to the low number of DEG. Nevertheless, disease and biological functions found with IPA in LW and CR after restriction were mainly related to the immune response (quantity of lymphocytes and T-lymphocytes, immune response of cells). Interestingly, only 3 disease and biological functions were activated after restriction in both breeds, which were “organismal death”, “anemia” and “polycythemia”, suggesting that feed restriction may also trigger genes associated with organismal death and blood defects. The canonical pathway comparison between breeds led to only one common pathway with a z-score < 2 in both LW and CR, which was synaptogenesis. Chronic stress exposure in rats and non-human primates have been shown to induce atrophy of dendrites and decreased glia and neurogenesis in the adult hippocampus [27], [28]. The mechanisms that control food intake also involve communication between gut, adipose tissue and the central nervous system through hormones and peptides circulating in the blood. We could therefore hypothesize that feed restriction generates stressful signals that may affect synaptogenesis.

The Top 5 canonical pathways found with in IPA in the two breeds after restriction did not overlap, suggesting differential response to feed restriction between breeds. In LW, several DEGs in the Top3 enriched pathways found in IPA encodes for Heat Shock Protein (HSPs): DNAJA1, DNAJC17, DNAJC9, HSP90AA1, HSPA12B. HSPs are highly conserved proteins playing an essential role in the cellular stress response [29]. The expression of HSP could be linked to the fact that the present experiment takes place in a tropical climate, with a mean temperature of 25.5°C, which is above growing pig thermoneutral temperature [2]. However, the differential expression of HSP was found comparing RF and NF after restriction, indicating that the response observed is related to the feed diet. Proteomic studies on short-term heat stress (12h) using pair-feeding controls showed that pigs with a reduced plane of nutrition in thermoneutral conditions had increased HSP70 [30]. HSP are also part of the common over-represented pathways Aldosterone Signaling in Epithelial Cells, Protein ubiquitination pathway and Unfolded Protein Signaling. Genes encoding for HSPs and involved in the aldosterone pathway have been identified as over-expressed in the liver and duodenum of pigs with low FE compared to high FE pigs [31]. Interestingly here, upregulation of HSP after feed restriction is detected in LW and not in CR, indicating that HSP are not triggered upon feed restriction in that breed. This evidence suggest that LW have higher stress response than CR, which is supported by the performance results obtained and previous studies comparing LW and CR [32]. In line with these results, a study comparing HSP90 mRNA expression levels after heat stress in peripheral blood mononuclear cells of LW and CR found an increase of HSP90 mRNA expression in both breeds after 6h, but a significant decrease in CR pigs after 9h [33]. The authors suggested that the difference observed after 9h could be due to a reduced impact of heat stress on protein conformations in CR pigs.

After refeeding, the number of DEGs was higher in LW than CR, suggesting stronger response to refeeding in LW than CR. In LW, the KEGG pathways identified after refeeding were related to the immune response but also to thermogenesis. Thermogenesis could be triggered during refeeding due to increased feed intake compared to the restriction period, which may generate increased metabolic heat [34]. The immune response has also been shown to be triggered upon refeeding in beef cattle jejunum transcriptomic profile and could allow more dietary derived energy to be partitioned towards growth during re-alimentation [18]. However, despite the greater number of DEGs found in LW than CR after refeeding, the difference in terms of performance between the 2 breeds after refeeding were not significant (data not shown). In none of the breeds do we observe compensatory growth, i. e. a period of accelerated growth following periods of feed restriction, during refeeding. Compensatory growth in pigs depends on the onset, severity and duration of the restriction period and the onset and duration of refeeding [35]. In the present study, despite a long period of feed restriction and refeeding, the severity of the feed restriction was probably not sufficient to induce compensatory growth. Consistent with this result, pathways and disease and biological functions enrichment in IPA after refeeding led to lower z-score than after restriction, suggesting lower response for both breeds after refeeding than after restriction.

5. Conclusions

In conclusion, the present study indicates that blood transcriptomics can be a useful tool to study differential genetic response to feed restriction in a dynamic way throughout the different periods of stress of the animal life. In both breeds, major transcriptional changes after restriction and refeeding were related to the immune

response. Nevertheless, the transcriptional response to feed restriction and refeeding between CR and LW was contrasted both in terms of number of DEGS and enriched pathways. CR demonstrated a stronger transcriptional response to feed restriction whereas LW had a stronger response to refeeding. Most differences in the transcriptional response to feed restriction between CR and LW were related to cell stress response, whereas after refeeding, differences were linked to thermogenesis, metabolic pathways and cell proliferation. Additional research on local breeds and potential structural variants that could increase the transcriptional response to feed restriction while maintaining performance would contribute to deepening our understanding of post-absorptive metabolism differences between breeds.

Acknowledgements

The authors gratefully thank all the members of staff and students who contributed to the project, especially K. Benony, B. Bocage, M. Bructer, M. Giorgi et F. Silou from the experimental unit INRAE-PTEA and J. Hira for RNA extraction. The financial support of EU-funds (FEDER, FSE, FEADER) and the Region Guadeloupe (including the AGROECODIV project) are gratefully acknowledged. We are also grateful to the Genotoul bioinformatics platform Toulouse Occitanie (Bioinfo Genotoul, <https://doi.org/10.15454/1.5572369328961167E12>) for providing help computing and storing the resources.

References

- [1] H. Laevens, F. Koenen, H. Deluyker, and A. De Kruif, 'Experimental infection of slaughter pigs with classical swine fever virus: Transmission of the virus, course of the disease and antibody response', *Vet. Rec.*, vol. 145, no. 9, pp. 243–248, 1999, doi: 10.1136/vr.145.9.243.
- [2] D. Renaudeau, J. L. Gourdine, and N. R. St-Pierre, 'A meta - analysis of the effects of high ambient temperature on growth performance of growing - finishing pigs', *J Anim Sci*, vol. 89, pp. 2220–2230, 2011, doi: 10.2527/jas.2010-3329.
- [3] C. M. Nyachoti, R. T. Zijlstra, C. F. M. de Lange, and J. F. Patience, 'Voluntary feed intake in growing-finishing pigs: A review of the main determining factors and potential approaches for accurate predictions', *Can. J. Anim. Sci.*, vol. 84, no. 4, pp. 549–566, 2004, doi: 10.4141/A04-001.
- [4] R. A. Donker, L. A. Den Hartog, E. W. Brascamp, J. W. M. Merks, G. J. Noordewier, and G. A. J. Buiting, 'Restriction of feed intake to optimize the overall performance and composition of pigs', *Livest. Prod. Sci.*, vol. 15, no. 4, pp. 353–365, 1986, doi: 10.1016/0301-6226(86)90075-8.
- [5] M. G. Hogberg and D. R. Zimmerman, 'Compensatory Responses to Dietary Protein, Length of Starter Period and Strain of Pig', *J. Anim. Sci.*, vol. 47, no. 4, pp. 893–899, Oct. 1978, doi: 10.2527/jas1978.474893x.
- [6] N. Pouillet *et al.*, 'Effect of feed restriction and refeeding on performance and metabolism of European and Caribbean growing pigs in a tropical climate', *Sci. Rep.*, vol. 9, no. 1, p. 4878, 2019, doi: 10.1038/s41598-019-41145-w.
- [7] P. J. Boettcher *et al.*, 'Genetic resources and genomics for adaptation of livestock to climate change', *Front. Genet.*, vol. 5, p. 461, Jan. 2015, doi: 10.3389/fgene.2014.00461.
- [8] J. L. Gourdine, J. P. Bidanel, J. Noblet, and D. Renaudeau, 'Effects of breed and season on performance of lactating sows in a tropical humid climate1', *J. Anim. Sci.*, vol. 84, no. 2, pp. 360–369, Feb. 2006, doi: 10.2527/2006.842360x.
- [9] J. L. Gourdine, J. P. Bidanel, J. Noblet, and D. Renaudeau, 'Effects of season and breed on the feeding behavior of multiparous lactating sows in a tropical humid climate', *J. Anim. Sci.*, vol. 84, no. 2, pp. 469–480, Feb. 2006, doi: 10.2527/2006.842469X.
- [10] M. Jégou, F. Gondret, A. Vincent, C. Tréfeu, H. Gilbert, and I. Louveau, 'Whole Blood Transcriptomics Is Relevant to Identify Molecular Changes in Response to Genetic Selection for Feed Efficiency and Nutritional Status in the Pig', *PLOS ONE*, vol. 11, no. 1, p. e0146550, Jan. 2016, doi: 10.1371/journal.pone.0146550.
- [11] A. Dobin *et al.*, 'STAR: Ultrafast universal RNA-seq aligner', *Bioinformatics*, vol. 29, no. 1, pp. 15–21, 2013, doi: 10.1093/bioinformatics/bts635.
- [12] S. Anders, P. T. Pyl, and W. Huber, 'HTSeq-A Python framework to work with high-throughput sequencing data', *Bioinformatics*, vol. 31, no. 2, pp. 166–169, 2015, doi: 10.1093/bioinformatics/btu638.
- [13] M. I. Love, W. Huber, and S. Anders, 'Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2', *Genome Biol.*, vol. 15, no. 12, pp. 1–21, 2014, doi: 10.1186/s13059-014-0550-8.
- [14] A. Krämer, J. Green, J. Pollard Jr, and S. Tugendreich, 'Causal analysis approaches in Ingenuity Pathway Analysis', *Bioinformatics*, vol. 30, no. 4, pp. 523–530, Feb. 2014, doi: 10.1093/bioinformatics/btt703.

- [15] S. X. Ge, D. Jung, and R. Yao, 'ShinyGO: a graphical gene-set enrichment tool for animals and plants', *Bioinformatics*, vol. 36, no. 8, pp. 2628–2629, Apr. 2020, doi: 10.1093/BIOINFORMATICS/BTZ931.
- [16] W. Luo and C. Brouwer, 'Pathview: An R/Bioconductor package for pathway-based data integration and visualization', *Bioinformatics*, vol. 29, no. 14, pp. 1830–1831, 2013, doi: 10.1093/bioinformatics/btt285.
- [17] M. Kanehisa, M. Furumichi, Y. Sato, M. Ishiguro-Watanabe, and M. Tanabe, 'KEGG: Integrating viruses and cellular organisms', *Nucleic Acids Res.*, vol. 49, no. D1, pp. D545–D551, 2021, doi: 10.1093/nar/gkaa970.
- [18] K. Keogh, S. M. Waters, P. Cormican, A. K. Kelly, and D. A. Kenny, 'Effect of dietary restriction and subsequent re-alimentation on the transcriptional profile of bovine jejunal epithelium', *PLOS ONE*, vol. 13, no. 3, p. e0194445, Mar. 2018, doi: 10.1371/JOURNAL.PONE.0194445.
- [19] K. Keogh, D. A. Kenny, P. Cormican, A. K. Kelly, and S. M. Waters, 'Effect of dietary restriction and subsequent re-alimentation on the transcriptional profile of hepatic tissue in cattle', *BMC Genomics*, vol. 17, no. 244, Mar. 2016, doi: 10.1186/S12864-016-2578-5.
- [20] N. Da Costa, C. McGillivray, Q. Bai, J. D. Wood, G. Evans, and K. C. Chang, 'Restriction of dietary energy and protein induces molecular changes in young porcine skeletal muscles', *J. Nutr.*, vol. 134, no. 9, pp. 2191–2199, 2004, doi: 10.1093/jn/134.9.2191.
- [21] C.-C. Liew, J. Ma, H.-C. Tang, R. Zheng, and A. A. Dempsey, 'The peripheral blood transcriptome dynamically reflects system wide biology: a potential diagnostic tool', *J. Lab. Clin. Med.*, vol. 147, no. 3, pp. 126–132, Mar. 2006, doi: 10.1016/j.lab.2005.10.005.
- [22] N. Mach *et al.*, 'The peripheral blood transcriptome reflects variations in immunity traits in swine: Towards the identification of biomarkers', *BMC Genomics*, vol. 14, no. 1, 2013, doi: 10.1186/1471-2164-14-894.
- [23] O. Lamas, J. A. Martínez, and A. Marti, 'Energy restriction restores the impaired immune response in overweight (cafeteria) rats', *J. Nutr. Biochem.*, vol. 15, no. 7, pp. 418–425, Jul. 2004, doi: 10.1016/j.jnutbio.2004.02.003.
- [24] F. Wasinski *et al.*, 'Exercise and Caloric Restriction Alter the Immune System of Mice Submitted to a High-Fat Diet', *Mediators Inflamm.*, vol. 2013, p. e395672, Mar. 2013, doi: 10.1155/2013/395672.
- [25] T. Ahmed, S. K. Das, J. K. Golden, E. Saltzman, S. B. Roberts, and S. N. Meydani, 'Calorie Restriction Enhances T-Cell-Mediated Immune Response in Adult Overweight Men and Women', *J. Gerontol. Ser. A*, vol. 64A, no. 11, pp. 1107–1113, Nov. 2009, doi: 10.1093/gerona/glp101.
- [26] Elsasser TH, Klasing KC, Filipov N, Thompson F., 'The metabolic consequence of stress: targets for stress and priorities of nutrient use', in *The biology of animal stress*, New York: CABI Publishing, 2000, pp. 77–110.
- [27] M. Banasr, G. W. Valentine, X.-Y. Li, S. L. Gourley, J. R. Taylor, and R. S. Duman, 'Chronic Unpredictable Stress Decreases Cell Proliferation in the Cerebral Cortex of the Adult Rat', *Biol. Psychiatry*, vol. 62, no. 5, pp. 496–504, Sep. 2007, doi: 10.1016/j.biopsych.2007.02.006.
- [28] R. M. Shansky and J. H. Morrison, 'Stress-induced dendritic remodeling in the medial prefrontal cortex: Effects of circuit, hormones and rest', *Brain Res.*, vol. 1293, pp. 108–113, Oct. 2009, doi: 10.1016/j.brainres.2009.03.062.
- [29] R. A. Stetler *et al.*, 'Heat shock proteins: Cellular and molecular mechanisms in the central nervous system', *Prog. Neurobiol.*, vol. 92, no. 2, pp. 184–211, Oct. 2010, doi: 10.1016/j.pneurobio.2010.05.002.
- [30] S. C. Pearce, S. M. Lonergan, E. Huff-Lonergan, L. H. Baumgard, and N. K. Gabler, 'Acute Heat Stress and Reduced Nutrient Intake Alter Intestinal Proteomic Profile and Gene Expression in Pigs', *PLOS ONE*, vol. 10, no. 11, p. e0143099, Nov. 2015, doi: 10.1371/journal.pone.0143099.
- [31] Y. Ramayo-Caldas *et al.*, 'Integrative approach using liver and duodenum RNA-Seq data identifies candidate genes and pathways associated with feed efficiency in pigs', *Sci. Rep.*, vol. 8, no. 1, p. 558, 2018, doi: 10.1038/s41598-017-19072-5.
- [32] J. Gourdine, W. M. Rauw, H. Gilbert, and N. Poulet, 'The Genetics of Thermoregulation in Pigs : A Review', *Front. Vet. Sci.*, vol. 8, no. December, 2021, doi: 10.3389/fvets.2021.770480.
- [33] J.-C. Bambou, J.-L. Gourdine, R. Grondin, N. Vachery, and D. Renaudeau, 'Effect of heat challenge on peripheral blood mononuclear cell viability: comparison of a tropical and temperate pig breed', *Trop. Anim. Health Prod.*, vol. 43, no. 8, pp. 1535–1541, Dec. 2011, doi: 10.1007/s11250-011-9838-9.
- [34] L. J. Koong, J. A. Nienaber, H. J. Mersmann, and R. L. Hruska, 'Effects of Plane of Nutrition on Organ Size and Fasting Heat Production in Genetically Obese and Lean Pigs', *J Nutr.*, vol. 113, pp. 1626–1631, 1983.
- [35] P. A. Lovatto, D. Sauvant, J. Noblet, S. Dubois, and J. Van Milgen, 'Effects of feed restriction and subsequent refeeding on energy utilization in growing pigs', *J Anim Sci*, vol. 84, pp. 3329–3336, 2006, doi: 10.2527/jas.2006-048.

Drug effects in gene regulation: how multi-omics integration can benefit gene therapy design

Emeline Cherchame¹, Thomas Gareau¹, Beáta György¹, Justine Guégan¹

¹ Paris Brain Institute, Data Analysis Core platform, *Hôpital de la Pitié-Salpêtrière*, 75013, Paris, France

Corresponding Author: emeline.cherchame@icm-institute.org

Abstract

Data Analysis Core (DAC) is part of Paris Brain Institute (ICM), a research center dedicated to neuroscience. DAC provides expertise in processing, integrating and analyzing complex data. Omics can provide several levels of information related to regulation of gene expression and integration of multi-omics data may widen the perspective on particular biological processes. Here we present the relevance of a multi-omics project designed to assess the effect of drugs on the activation of specific metabolic pathways of interest for gene therapy.

Three drugs and their 2 control vehicles were applied on primary cells of 5 healthy donors. After 5h, total RNA and small RNA were purified to launch RNA-seq, miRNA-seq and ATAC-seq. RNA-seq pipeline was performed on Illumina DRAGEN bio-IT Platform. The pipeline to quantify miRNAs was implemented as proposed by Potla et al., 2021. Differential gene expression analysis was performed using edgeR. A batch effect was observed on donors in all technologies. Correction of the batch effect was implemented using the limma RemoveBatchEffect function. ATAC-seq pipeline was implemented with Snakemake, according to Encode recommendations. Peak count matrix was generated using FeatureCounts based on Superpeaks approach. Differentially accessible regions analysis was performed using limma, with covariate on donors. Integration of RNAseq and miRNAseq results was performed according to Yan et al., 2019 methodology. Pearson's correlation test was conducted to examine pairwise correlations between deregulated-miRNAs and deregulated genes. Anti-correlation was visualized as a network using Cytoscape. miRNA target genes were predicted using program multiMiR (Ru et al., 2014). Integration RNA/ATAC was performed by intersected ATAC-seq DARs and RNA-seq DEGs results, based on overlaps between nearest TSS of peak and deregulated genes, respectively.

RNA-seq data showed drugs A and C have effect on the transcriptome regulation, drug B as no effect. ATAC-seq DARs analysis showed several chromatin rearrangements induce by drug A. No chromatin rearrangement was observed with drug B and C. A motif enrichment analysis on peak regions was performed by HOMER findMotifsGenome.pl function and MEME-SUITE tool. The target transcription factor motif has been identified. The integration of multi-omics data allowed to confirm the strategy with drugs to modulate gene expression.

Keywords

RNA-seq, ATAC-seq, miRNA, multi-omics data integration, transcriptomic, epigenomic, motifs enrichment, gene therapy

References

1. Pratibha Potla, Shabana Amanda Ali, Mohit Kapoor, (2021) A bioinformatics approach to microRNA-sequencing analysis, *Osteoarthritis and Cartilage Open*, Volume 3, Issue 1, March 2021, 100131, <https://doi.org/10.1016/j.ocarto.2020.100131>.
2. Yao Y, Jiang C, Wang F, Yan H, Long D, Zhao J, Wang J, Zhang C, Li Y, Tian X, Wang QK, Wu G, Zhang Z. Integrative Analysis of miRNA and mRNA Expression Profiles Associated With Human Atrial Aging. *Front Physiol.* 2019 Sep 19;10:1226. doi: 10.3389/fphys.2019.01226. PMID: 31607954; PMCID: PMC6761282.
3. Ru Y, Kechris KJ, Tabakoff B, Hoffman P, Radcliffe RA, Bowler R, Mahaffey S, Rossi S, Calin GA, Bemis L, Theodorescu D. The multiMiR R package and database: integration of microRNA-target interactions along with their disease and drug associations. *Nucleic Acids Res.* 2014;42(17):e133. doi: 10.1093/nar/gku631. Epub 2014 Jul 24. PMID: 25063298; PMCID: PMC4176155

AnnotSV and knotAnnotSV: a webserver for human structural variations annotations and analysis

Véronique GEOFFROY¹², Jean-Baptiste LAMOUCHE²³, Thomas GUIGNARD⁴, Samuel NICAISE³, Arnaud KRESS⁵, Sophie SCHEIDECKER²⁶, Antony LE BECHEC³ and Jean MULLER²³⁶

¹ Université de Brest, INSERM, EFS, UMR 1078, GGB, F-29200 Brest, France

² Laboratoire de Génétique Médicale, UMR 1112, INSERM, IGMA, Université de Strasbourg, Strasbourg, France

³ Unité Fonctionnelle de Bioinformatique Médicale appliquée au diagnostic (UF7363), Hôpitaux Universitaires de Strasbourg, Strasbourg, France

⁴ Unité de Génétique Chromosomique, CHU Montpellier, France

⁵ Complex Systems and Translational Bioinformatics, ICube, UMR 7357, University of Strasbourg, CNRS, FMTS, Strasbourg, France

⁶ Laboratoires de Diagnostic Génétique, IGMA, Hôpitaux Universitaires de Strasbourg, Strasbourg, France

Corresponding Author: veronique.geoffroy@inserm.fr

Website: <https://lbgi.fr/AnnotSV/>

Abstract

Thanks to the large interest in human medical genomics and the facilitated access to pangenomic analysis, Human geneticists have generated large amount of genomic data. Indeed, millions of small variants (SNV/Indel) and thousands of structural variations (SV) are identified from next-generation sequencing, array-based techniques but also now optical genome mapping. To help analysing human pathogenic SV, we present the new version of our webserver dedicated to their annotation, ranking and visualization available at the following address: <https://www.lbgi.fr/AnnotSV/>.

Since the first AnnotSV version [1] and the webserver publication [2], we have continuously provided updated annotations, updated methods and visualizations. Novel annotations include miRNA data annotations, cytoband definitions, consistent terminology from GenCC, Benign SV from Children's Mercy Research Institute WGS. The phenotype driven module has been updated considering the latest Exomiser version available. The automatic ranking based on the ACMG/ClinGen recommendations [3] has been updated to add 8 supplementary subsections and enhance the ranking precision. In parallel, the ranking description has been improved.

Finally, the webserver allows the visualization of SV using 3 different methods: An interactive web page, an interactive circos view and a handy spreadsheet mode. Input files include either standard bed or vcf files. Output can be either visualized in a web browser directly or downloaded as separated files including the newly proposed vcf output file (Figure 1).

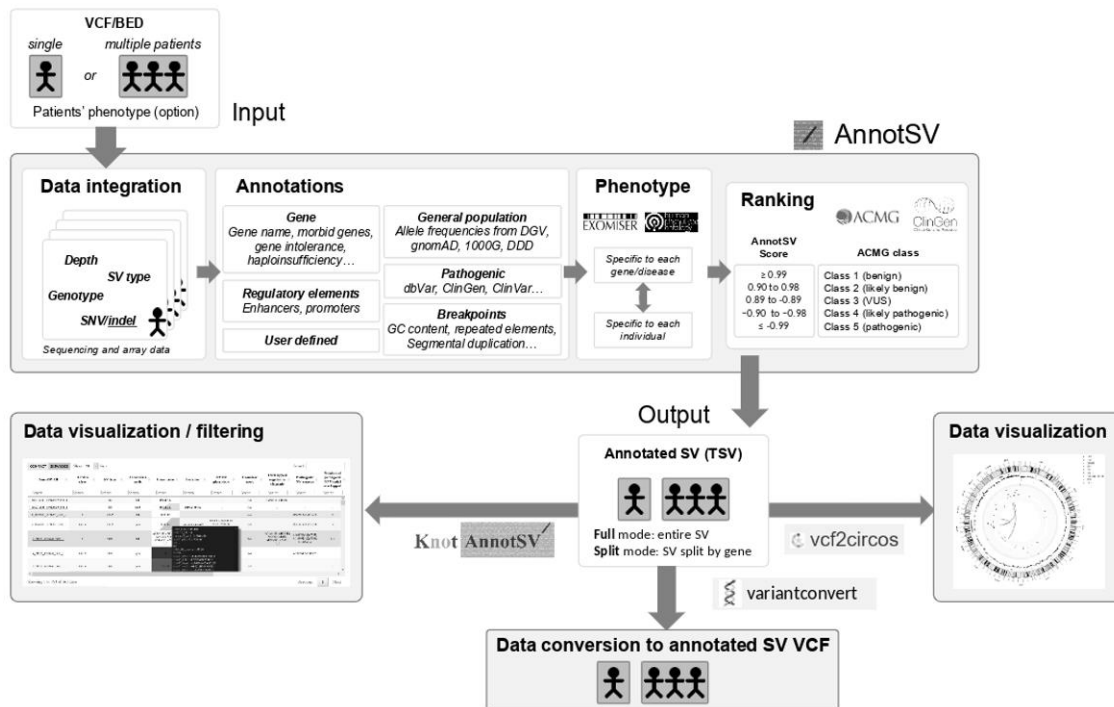


Figure 1. AnnotSV and knotAnnotSV workflow.

We are convinced that this comprehensive online SV annotation and interpretation tool is of great interest for the biologists interested in human genomics.

This website is free and open to all users and there is no login requirement.

Keywords Structural variations, Annotation, Ranking, Visualisation, Human genome

References

1. Geoffroy Véronique. AnnotSV: An integrated tool for Structural Variations annotation. *Bioinformatics*, 2018. doi: [10.1093/bioinformatics/bty304](https://doi.org/10.1093/bioinformatics/bty304)
2. Geoffroy Véronique and Guignard Thomas. AnnotSV and knotAnnotSV: a webserver for human structural variations annotations and analysis. *Nucleic Acid Research.*, 2021. doi: [10.1093/nar/gkab402](https://doi.org/10.1093/nar/gkab402)
3. Erin Rooney Riggs. Technical Standards for the Interpretation and Reporting of Constitutional Copy-Number Variants: A Joint Consensus Recommendation of the American College of Medical Genetics and Genomics (ACMG) and the Clinical Genome Resource (ClinGen). *Genetics in Medicine*, 2020. doi : [10.1038/s41436-019-0686-8](https://doi.org/10.1038/s41436-019-0686-8)

Cirscan: a shiny application to identify differentially active sponge mechanisms and visualize circRNA-miRNA-mRNA networks.

Rose-Marie Fraboulet¹, Yanis Si Ahmed¹, Marc Aubry², Sébastien Corre¹,
Marie-Dominique Galibert^{1,3}, Yuna Blum^{1*}

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

² INSERM, OSS (Oncogenesis Stress Signaling), UMR_S 1242, CLCC Eugene Marquis, Univ Rennes 1, 35000, Rennes, France.

³ Department of Molecular Genetics and Genomics, Hospital University of Rennes (CHU Rennes), F-35000 Rennes, France.

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

Abstract *Non-coding RNAs and their biological functions are still incompletely understood, while representing a large part of the human transcriptome and having an important role in cancer. Among these, circular RNAs (circRNAs) have recently been discovered for their microRNA (miRNA) sponge function, which allows them to modulate the expression of miRNA target genes: they take on the role of competitive endogenous RNAs (ceRNAs). Their closed-loop structure makes them highly stable and their miRNA binding capacity seems much more powerful than that of any other ceRNAs, leading to their super-sponge naming. Today, few ceRNA prediction computational tools have been published and most of them do not consider ce-circRNAs. Moreover, the few studies focusing on circRNA-miRNA-mRNA networks have not developed a tool that automates the search for ceRNAs from user's transcriptomic data. In this study, we present an interactive Shiny web application, called Cirscan for CIRcular RNA Sponge CANDidates. Cirscan automatically infers circRNA-miRNA-mRNA networks from human multi-level transcript expression data (including circRNAs, mRNAs, and optionally miRNAs in the case of cancer study) from two biological conditions (e.g. tumor versus normal conditions) in order to identify potential sponge mechanisms, active in a specific condition. Cirscan calculates a global sponge score for each miRNA-circRNA interaction, integrating multiple criteria based on interaction reliability and on RNA expression level using the TOPSIS method. The ranking of this score provides the most relevant circRNA candidates being miRNA sponges. Finally, the user is able to visualize the top ranked sponge mechanisms as graphs, where nodes correspond to the different types of RNAs and edges to the miRNA-target interactions. In order to help the user in the biological interpretation of the visualized network, an enrichment analysis of the biological functions is performed. We applied Cirscan on a public multi-level microarray transcript expression dataset from colorectal cancer (CRC) and retrieved previously described sponge mechanisms as well as novel circRNA sponge candidates.*

Keywords circular RNAs, sponge mechanism, transcriptomic data, regulation network, cancer

1. Introduction

Non-coding RNAs (ncRNAs), genome sequences not translated into proteins, and their biological functions are still incompletely understood, because they were characterized as junk transcriptional products. Recently, it has been demonstrated that ncRNAs represent a large part of the human transcriptome and have important functional roles, particularly in cancers [1]. Advances in sequencing technologies have led to the identification of different types of ncRNAs in the cell, such as long non-coding RNAs and microRNAs (miRNAs) [2]. In this study, we are interested in circular RNAs (circRNAs), recently discovered and known for their miRNA sponge function [3,4]. CircRNAs can indirectly regulate the expression of genes involved in cell plasticity by sequestering miRNA(s) and can

play a role of a competitor, called competitive endogenous RNA (ceRNA) [5–7]. One of the most well-known ce-circRNA is ciRS-7 (CDR1as), mainly expressed in the brain, which acts as a regulator of the miRNA-7 with more than 70 binding sites and leads to increased expression levels of miRNA-7 targets [8]. Thanks to their closed-loop structure formed by reverse-splicing, circRNAs are very stable and their binding capacity to bind to miRNAs seems to be stronger than that of any other ceRNA, leading to their super-sponge naming [9]. These interactions between coding and non-coding RNAs can be represented as circRNA-miRNA-mRNA interaction networks.

Today, few computational tools have been developed for the identification of ceRNA but are not adapted to the search for ce-circRNA [10,11]. Moreover, the studies focusing on circRNA-miRNA-mRNA networks have not developed a tool that automates the search for ce-circRNA from its own transcriptomic data [12–18]. Recently, two circRNA functional annotation databases predicting sponge mechanisms in several tissues have been developed [19,20], but it is not possible to query one's own dataset and to identify sponge mechanisms specific to a biological condition. Here, we present an interactive Shiny web application, called Cirscan for CIRcular RNA Sponge CANDidates. The aim of this tool is to automatically infer circRNA-miRNA-mRNA networks from human multi-level transcript expression data (including circRNAs, mRNAs, and optionally miRNAs in the case of cancer study) from two biological conditions (e.g. tumor versus normal conditions) and to identify sponge mechanisms, active in a specific condition. From the user's expression data and an in-house database of highly reliable prediction interactions, Cirscan is able to construct an interaction score matrix including multiple criteria based on interaction reliability (e.g. affinity of the interaction, enrichment of binding sites for a given miRNA) and on RNA expression (e.g. minimum level of expression, co-expression between circRNA and mRNA targets for a given miRNA). To optimize the search for effective sponge mechanism, a global sponge score is calculated using the different criteria for each miRNA-circRNA predicted interaction. The ranking of this global sponge score provides the most relevant circRNA candidates being miRNA sponges. Finally, the user is able to visualize the top ranked sponge mechanisms as graphs, where nodes correspond to the different types of RNAs and edges to the miRNA-target interactions. In order to help the user in the biological interpretation of the results and to better understand the functional impact of the selected sponge mechanism, an enrichment analysis of the biological functions is performed by Cirscan for each visualized network.

We applied our tool on a public multi-level transcript expression dataset from two conditions: colorectal cancer samples and normal adjacent samples [21]. We showed that Cirscan was able to retrieve previously described sponge mechanisms as well as novel circRNA sponge candidates [22].

In summary, Cirscan provides a user-friendly tool to identify and visualize circRNA-miRNA-mRNA networks with potential sponge mechanisms from user's human multi-level transcript expression data, and may provide potential novel biomarkers for the development of RNA targeted therapies [23].

Cirscan shiny app is freely available on Gitlab at the following link: https://gitlab.com/geobioinfo/cirscan_Rshiny.

2. Software description

2.1. Overview of the tool

Cirscan infers circRNA-miRNA-mRNA interactions from coding and non-coding transcriptomic data (Fig. 1A) and identifies circRNA candidates that can act as miRNA sponges (ce-circRNA). The tool comprises the following main steps: 1) Construction of an interaction score matrix including different criteria: criteria of interaction reliability and criteria of effectiveness of a sponge mechanism as defined in the following sections (Fig. 1B. Step 1). 2) Calculation of a global sponge score for each circRNA-miRNA interaction by integrating the different criteria using the TOPSIS method [24] [see more details in the section 2.4.3] (Fig. 1B. Step 2). Cirscan outputs the top ranked circRNA sponge candidates based on the ranked global sponge scores. It also enables the visualization of the identified sponge mechanisms as networks and biological enrichment analysis of the mRNAs of the networks of interest (Fig. 1C). A video tutorial and an example dataset are available to guide the user in the different steps on the application.

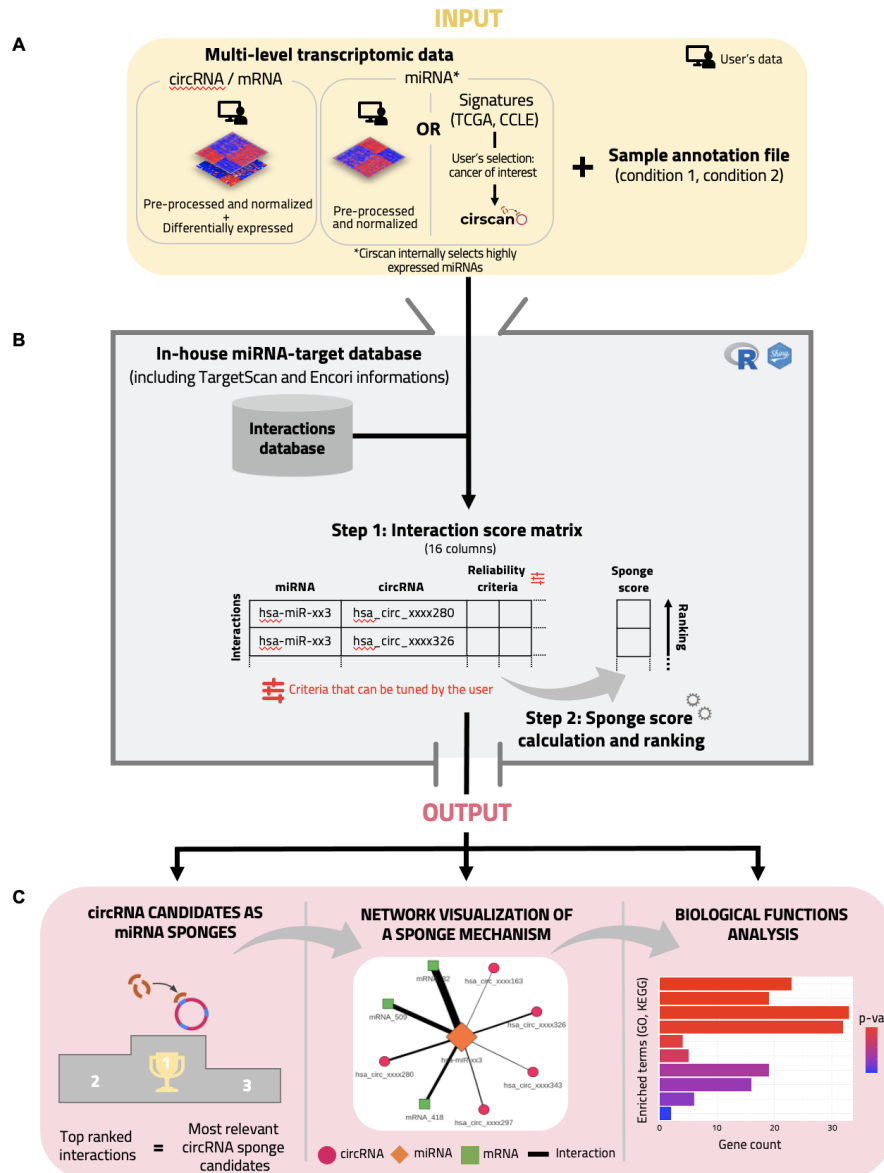


Fig. 1 Cirscan workflow. (A) Input files provided by the user. (B) Integration of multiple criteria for the calculation of a predicted sponge score for each circRNA-miRNA pair. (B. Step 1) Construction of an RNA interaction matrix including multiple criteria based on interaction reliability and on RNA expression. (B. Step 2) Calculation of the sponge score. (C) Outputs produced by Cirscan.

2.2. Effective sponge mechanisms prerequisites

Prerequisites based on both sequence and expression information have been put forward for the identification of circRNAs acting as miRNA sponges: 1) A miRNA must be sufficiently expressed whatever the biological conditions to have a functional impact on its targets and must target at least one circRNA and one mRNA, with a strong affinity. Importantly we do not assume a differential expression between both conditions due to a sponge mechanism: we assume that a sponge mechanism involving a circRNA only affects the bioavailability of the miRNA but does not affect the miRNA expression. 2) The mRNAs and circRNAs that are targets of the miRNA must be co-expressed, i.e differentially expressed between the two conditions in the same direction. Indeed, it is expected that the increased presence of a circRNA that can sponge a miRNA will induce an over-expression of the mRNA targets of the miRNA, as the miRNA can no longer degrade them.

2.3. Input files

Cirscan requires two types of CSV files from the user that are depicted in Figure 1A. First, multi-level transcript expression data are required as input files, including circRNA, mRNA and optionally miRNA expression matrices. If the user does not have miRNA expression data, miRNA signatures for different types of cancer (respectively 32 and 20 from the TCGA consortium (<https://www.cancer.gov/tcga>) and CCLE database [25] (www.broadinstitute.org/ccle)) are internally available in Cirscan. There is no constraint concerning the technology used to obtain the expression matrices (array or sequencing), which in return requires the user to properly normalize his data beforehand. Each expression matrix needs to be normalized (e.g. RMA for expression array, TPM for RNA-seq data) and log-transformed, and a minimum of 3 samples per condition is required to ensure reliability of the downstream analysis. Based on the sponge mechanism prerequisites, circRNA and mRNA expression matrices should be restricted to the differentially expressed circRNAs and mRNAs prior to the use of Cirscan. As mentioned above, this restriction is not relevant for miRNAs and no filter is required. Cirscan will automatically keep miRNAs that are sufficiently expressed considering the following assumptions: (1) the sum of the expression of a given miRNA in all samples must be non-zero, (2) for a given miRNA, its expression must be higher than a defined cutoff in more than 90% of the samples, the cutoff being the quartile Q1 of the overall miRNA expression distribution. This selection was also made for the establishment of the miRNA signatures from the TCGA and CCLE datasets.

CircRNA identifiers must correspond to the human circRNAs referenced in circBase (“hsa_circ_XXXXXXX”) [26], human miRNAs from miRBase (“hsa-miR-xxx” or “hsa-let-xxx”) [27], and mRNAs must be referenced as gene symbols. Furthermore, the user must provide another CSV file corresponding to the sample annotation file. This file must contain two columns: “samples” with the sample names referenced in the RNA expression matrices and “conditions” with the corresponding condition it refers to (“condition_1”, “condition_2”), knowing that the first condition refers to the control condition. The file format required by Cirscan is detailed in the Home panel of the Shiny application and in the example dataset provided in the application.

2.4. Identification of sponge mechanisms

2.4.1. In-house miRNA-target interaction databases

A miRNA-target interaction is established between a region of the miRNA, called “seed sequence” and a complementarity sequence called “miRNA response element” (MRE) on the target (circRNA or mRNA). An in-house database of interactions between miRNAs and putative mRNA or circRNA targets was constructed using the TargetScan interaction prediction tool and information from experimentally validated interaction database (ENCORI [28]) (Fig. 1B).

TargetScan was used to predict miRNA-circRNA and miRNA-mRNA interactions, and to obtain an affinity score for each interaction, by taking into account the specificity of the RNA type considered (mRNA or circRNA). For miRNA-circRNA interactions, we used the context+ score available in the TargetScan v6 [29]. The context+ score is the sum of the contribution of six features (site-type, 3' pairing, local AU, position, target site abundance, seed-pairing stability). The feature 3' pairing was removed here because it is not relevant for circRNA as the pairing is not expected to be restricted to that region and can appear on the whole circRNA sequence. The lower the context+ score, the stronger the affinity between the miRNA and its target. To reduce the number of false positive predictions, we defined a cutoff on the context+ score, based on its distribution on experimentally validated interactions given by the ENCORI database. We defined as a cutoff the 95th percentile of the context+ score distribution restricted to the interactions supported by at least two CLIP-seq and degradome-seq experiments. For miRNA-mRNA interactions, we calculated the context++ score provided by TargetScan v8 [30]. Compared to the context+ score, the context++ score includes additional criteria specifically relevant for miRNA-mRNA interactions (e.g. 3' UTR length, ORF length, probability of conserved targeting between species), which are expected to reduce false positive predictions.

MiRNA-circRNA and miRNA-mRNA interactions were restricted to specific MRE binding sites: 7mer-m8 (exact match to positions 2-8 of the mature miRNA), 7mer-1a (exact match to positions 2-7 of the mature miRNA followed by an “A”), and 8mer-1a (exact match to positions 2-8 of the mature miRNA followed by an “A”).

2.4.2. Criteria of interaction reliability and sponge effectiveness

In order to identify potential sponge mechanisms involving circRNA, different scores of reliability are calculated. We defined as the affinity score between a miRNA and its targets $S_{Affinity}^{miRNA-target}$ the context score calculated by TargetScan that we rescaled between 0 and 1: the closer to 1, the stronger the interaction affinity. For each miRNA-target interaction, we also considered the number of MRE sites that we named the score $S_{MRE}^{miRNA-target}$. If an interaction involves multiple MREs for the same miRNA and target of interest, the interaction with the maximum $S_{Affinity}^{miRNA-target}$ is retained. To reduce the potential bias in the number of predicted MREs due to the length of the circRNA target sequence, the number of MREs for miRNA-circRNA interactions was normalized by the length of the circRNA sequence as described in the Cerina tool [19] and referenced here as the score $S_{MRE/kb}^{miRNA-target}$. In order to identify circRNAs that are significantly enriched for MRE binding sites of specific miRNAs, we used a binomial model conditioned on the total number of MRE for the given miRNA on each circRNA and the number of all other miRNA-MRE sites on the given circRNA as background, as proposed in the recent scanMiR tool [31] and defined as $S_{Enrichment}^{miRNA-target}$. A circRNA-miRNA interaction is kept if the Binomial enrichment test adjusted p-value (Benjamini Hochberg (BH)) is lower than 0.05.

In order to take into account the expression level information, different scores are calculated: the target expression fold change between conditions 1 and 2 ($S_{log(FC)}^{target}$), the median expression of each miRNA under all conditions ($S_{Expression}^{miRNA}$) and the highest mean expression of each target between the two conditions ($S_{Expression}^{target}$). Only networks with at least one circRNA and one mRNA with fold change values in the same direction are selected.

2.4.3. Calculation of a sponge score

For each miRNA-mRNA interaction, Cirscan calculates a global score defined as $SG_{Interaction}^{miRNA-mRNA}$ which integrates the different criteria described above ($S_{Affinity}^{miRNA-target}$, $S_{MRE}^{miRNA-target}$, $S_{log(FC)}^{target}$, $S_{Expression}^{miRNA}$ and $S_{Expression}^{target}$) using the TOPSIS method [24]. Briefly, the TOPSIS method is a multi-criteria decision analysis method which aims to determine the best alternative when different criteria need to be considered together. In our case, the best alternative is the miRNA-target interaction for which all criteria are maximized, e.g. highest affinity score and highly expressed miRNAs, circRNAs and mRNAs.

Then, a score for each miRNA-circRNA interaction is calculated and defined as a global sponge score, $SG_{Interaction}^{miRNA-circRNA}$. This score is also based on the TOPSIS method and uses the following criteria: $S_{Affinity}^{miRNA-target}$, $S_{MRE/kb}^{miRNA-target}$, $S_{Enrichment}^{miRNA-target}$, $S_{log(FC)}^{target}$, $S_{Expression}^{miRNA}$, $S_{Expression}^{target}$ and $S_{Enrichment}^{miRNA-miRNA}$, where $S_{Enrichment}^{miRNA-miRNA}$ corresponds to the enrichment score (NES) (fgsea R package [32]) calculated on the ranked global score $SG_{Interaction}^{miRNA-mRNA}$ of the mRNA targets defined above. In other words, for a given miRNA-circRNA interaction, if the mRNA targets of the miRNA have high global scores, this will contribute to increasing the sponge score. MiRNA-circRNA interactions are then ranked according to their sponge score, the top ranked interactions with the highest scores being the most relevant sponge circRNA candidates.

Finally, all this information is compiled into an “interaction score matrix” with 16 columns available and detailed in Cirscan (Fig. 1B), where $Rank_{miRNA-circRNA}$ is the rank of miRNA-circRNA interactions based on the global sponge score.

2.5. Visualization of the sponge mechanisms of interest

The identified circRNA-miRNA-mRNA networks involving potential sponge mechanisms can be visualized in the Network Visualization panel using the visNetwork package (v2.0.8, <https://github.com/datastorm-open/visNetwork>) (Fig. 1C). It is possible to visualize a network by

selecting an interaction from the “interaction score matrix” or by specifying the name of an RNA of interest: nodes correspond to the different types of RNAs (orange miRNAs, green mRNAs, and pink circRNAs) and edges to the miRNA-target interactions. The edges thickness is proportional to the global sponge score (the higher the value, the thicker the edge and the stronger the interaction affinity) and the RNA node of interest is larger than the others.

2.6. GO and KEGG enrichment analysis of mRNAs

In order to help the user in the biological interpretation of the visualized network, an enrichment of biological terms on the genes of the visualized network is performed using Enrichr tool [33] with KEGG (KEGG_2021_Human) and GO (GO_Biological_Process_2021) databases.

3. Application

We applied Cirscan on a public multi-level transcript expression dataset including mRNAs, miRNAs, and circRNAs of colorectal cancer (CRC) to infer sponge mechanisms involving circRNAs.

3.1. Pre-processing of input data

Microarray multi-level transcript expression data of 10 CRC samples and 10 normal adjacent samples were downloaded from the NCBI Gene Expression Omnibus database (accession number: GSE126095). Each dataset was imported into the RStudio (v1.4.1103) environment with R (v4.1.2) for pre-processing.

For transcripts belonging to the gene annotation, an expression average was applied. Quantile normalization and a log-transformation were applied to the mRNAs expression matrix. mRNAs microarray matrix was reduced to protein-coding genes, using the annotation file provided by the authors. Using the limma package [34] (v.3.50.1), 4,640 differentially expressed mRNAs with an adjusted p-value (BH) < 0.05 were selected. The circRNAs microarray matrix was also submitted to quantile normalization and a log-transformation. We selected 1,491 differentially expressed circRNAs with an adjusted p-value (BH) < 0.05 by using the limma package. Finally, the miRNAs microarray matrix was filtered to only include miRNA identifiers present in the annotation file, i.e. 2,055 miRNAs. Quantile normalization and log-transformation were also applied to the miRNAs expression matrix. These different pre-processed expression matrices were given as input to Cirscan.

In the sample annotation file, we considered the names of the 10 normal adjacent samples as “condition_1” and the 10 colorectal cancer samples as “condition_2”.

3.2. Results

Using the Cirscan tool on the multi-level transcript expression data from colorectal cancer described above, we identified 12,850 potential circRNA-miRNA sponge mechanisms involving 1,413 unique circRNAs. Among them, we identified 3 sponge mechanisms already described in the literature [22] (one sponge mechanism active in the normal condition and two in the tumor condition) significantly enriched in the top ranked sponge mechanisms (GSEA enrichment from fgsea R package [32], p-value = 0.00012). A similar result is observed using the TCGA colorectal cancer miRNA signature available internally in Cirscan (GSEA enrichment, p-value = 0.00015). Among the two subnetworks already described in the literature and specifically active in the tumor condition, we identified the hsa-circ-0001955:hsa-miR-145-5p sub-network, ranked 117 (in the top 1%) out of all identified mechanisms (Fig. 2A) and described as well by Ding *et al.* (2020) [22] in colorectal cancer. The biological term enrichment analysis of the mRNAs of this subnetwork revealed biological pathways associated with oncogenesis, such as Proteoglycans in cancer or MAPK signalling pathway (Fig. 2B). This result is consistent with the involvement of this sponge mechanism in the oncogenesis of colorectal cancer.

We next focused on the best ce-circRNA candidates having a sponge role in the tumor condition. Interestingly, within the top 10 ce-circRNA candidates, the first ce-circRNA has been already shown to have a sponge role in gastric or colorectal cancer in several studies (hsa_circ_0001658 [35–37]). Cirscan highlighted potential novel sponge mechanisms for this circRNA, involving for example the miRNA hsa-miR-665 (Fig. 2C) targeting genes associated to cell division and angiogenesis as

illustrated in Figure 2D (e.g. MAPK and VEGF signaling pathways). Four ce-circRNAs within the top 10 ce-circRNA candidates have been identified in other cancers as hepatocellular carcinoma and lung cancers (hsa_circ_0034326 [38], hsa_circ_0072088 [39–41], hsa_circ_0000517 [42,43], hsa_circ_0000326 [44]). The other best ce-circRNA candidates (hsa_circ_0087961, hsa_circ_0012283, hsa_circ_0007582, hsa_circ_0000254 and hsa_circ_0008720) have not yet been described in the literature as associated with cancer, making them interesting subjects for further experimental validations.

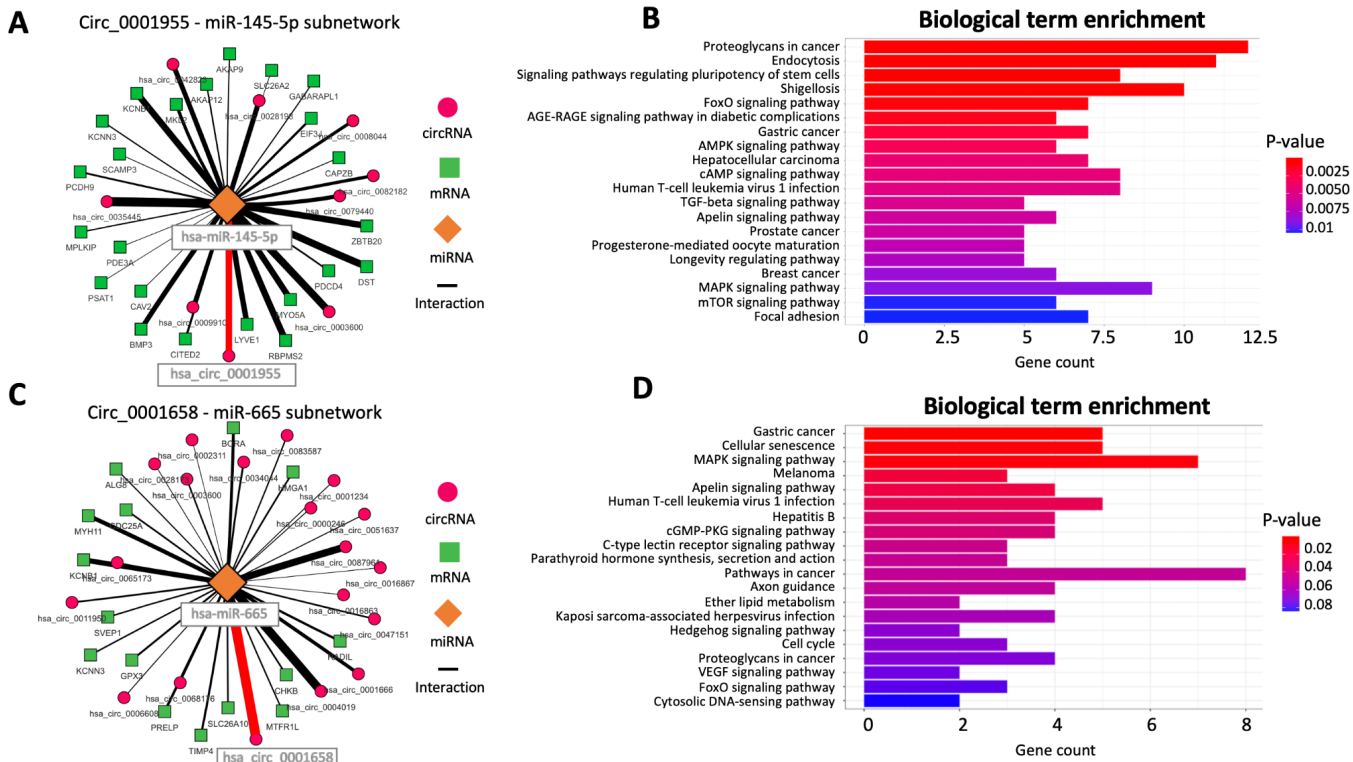


Fig. 2 Subnetworks identified by Cirscan using colorectal cancer data. (A) Circ_0001955-miR-145-5p subnetwork. The nodes represent the different RNA types (orange miRNAs, green mRNAs and pink circRNAs), and the edges represent miRNA-target interactions (targeting mRNAs or circRNAs). The red edges represent the interactions found in the literature and the thickness of the edges is proportional to the interaction score values. For a better visibility, only 10% of the targets are represented. (B) Top 20 mRNA-enriched pathways in the Circ_0001955-miR-145-5p subnetwork. (C) Circ_0001658-miR-665 subnetwork, with the same legend as panel A. (D) Top 20 mRNA-enriched pathways in the Circ_0001658-miR-665 subnetwork.

4. Conclusion and perspectives

Cirscan is a tool that takes as input any human multi-level transcript expression data (circRNAs, mRNAs, and optionally miRNAs) to identify and visualize condition-specific sponge mechanisms involving circRNAs. As shown using a public dataset from colorectal cancer tissues, Cirscan allows the identification of known and novel potential sponge mechanisms that may be further investigated and validated experimentally. This tool can be considered as a companion tool for biologists, facilitating their ability to prioritize sponge mechanisms for experimental validations. The mechanisms revealed by Cirscan could open new avenues for the development of novel RNA-targeted therapies. In particular, it would be possible to use antisense oligonucleotides, which can bind by complementarity to circRNA sequences of interest, to inhibit the sponge mechanisms active in a specific condition [23]. Finally, the framework established in this study could be extended to other species.

Acknowledgements

The authors would like to thank the Gene Expression and Oncogenesis team (CNRS UMR6290) for helpful discussion. We also acknowledge the GenOuest bioinformatics core facility (<https://www.genouest.org/>) for providing the computing infrastructure. This study received financial support from the Ligue National Contre le Cancer (LNCC) Départements du Grand-Ouest. RMF is a recipient of a doctoral fellowship from the Ligue National Contre le Cancer (LNCC) and YSA is a recipient of a doctoral fellowship from the LNCC Grand Ouest Départements du Grand-Ouest.

References

1. Anastasiadou E, Jacob LS, Slack FJ. Non-coding RNA networks in cancer. *Nat Rev Cancer*. 2018 Jan;18(1):5–18.
2. Esteller M. Non-coding RNAs in human disease. *Nat Rev Genet*. 2011 Dec;12(12):861–74.
3. Lacazette É, Diallo LH, Tatin F, Garmy-Susini B, Prats AC. L'ARN circulaire nous joue-t-il des tours ? *médecine/sciences*. 2020 Jan 1;36(1):38–43.
4. Liu J, Liu T, Wang X, He A. Circles reshaping the RNA world: from waste to treasure. *Mol Cancer*. 2017 Mar 9;16(1):58.
5. Zhang Z, Yang T, Xiao J. Circular RNAs: Promising Biomarkers for Human Diseases. *EBioMedicine*. 2018 Aug;34:267–74.
6. Hua X, Sun Y, Chen J, Wu Y, Sha J, Han S, Zhu X. Circular RNAs in drug resistant tumors. *Biomed Pharmacother Biomedecine Pharmacother*. 2019 Oct;118:109233.
7. Kristensen LS, Andersen MS, Stagsted LVW, Ebbesen KK, Hansen TB, Kjems J. The biogenesis, biology and characterization of circular RNAs. *Nat Rev Genet*. 2019 Nov;20(11):675–91.
8. Memczak S, Jens M, Elefsinioti A, Torti F, Krueger J, Rybak A, Maier L, Mackowiak SD, Gregersen LH, Munschauer M, Loewer A, Ziebold U, Landthaler M, Kocks C, le Noble F, Rajewsky N. Circular RNAs are a large class of animal RNAs with regulatory potency. *Nature*. 2013 Mar 21;495(7441):333–8.
9. Kristensen LS, Jakobsen T, Hager H, Kjems J. The emerging roles of circRNAs in cancer and oncology. *Nat Rev Clin Oncol*. 2022 Mar;19(3):188–206.
10. Chiu HS, Llobet-Navas D, Yang X, Chung WJ, Ambesi-Impiombato A, Iyer A, Kim HR, Seviour EG, Luo Z, Sehgal V, Moss T, Lu Y, Ram P, Silva J, Mills GB, Califano A, Sumazin P. Cupid: simultaneous reconstruction of microRNA-target and ceRNA networks. *Genome Res*. 2015 Feb;25(2):257–67.
11. Do D, Bozdogan S. Cancerin: A computational pipeline to infer cancer-associated ceRNA interaction networks. *PLOS Comput Biol*. 2018 juil;14(7):e1006318.
12. Yi Y, Liu Y, Wu W, Wu K, Zhang W. Reconstruction and analysis of circRNA-miRNA-mRNA network in the pathology of cervical cancer. *Oncol Rep*. 2019 Apr;41(4):2209–25.
13. Das A, Shyamal S, Sinha T, Mishra SS, Panda AC. Identification of Potential circRNA-microRNA-mRNA Regulatory Network in Skeletal Muscle. *Front Mol Biosci* [Internet]. 2021 [cited 2023 May 15];8. Available from: <https://www.frontiersin.org/articles/10.3389/fmolb.2021.762185>
14. Gong K, Yang K, Xie T, Luo Y, Guo H, Tan Z, Chen J, Wu Q, Gong Y, Wei L, Luo J, Yao Y, Yang Y, Xie L. Identification of circRNA-miRNA-mRNA regulatory network and its role in cardiac hypertrophy. *PLOS ONE*. 2023 Mar 23;18(3):e0279638.
15. Xiong D dan, Dang Y wu, Lin P, Wen D yue, He R quan, Luo D zhong, Feng Z bo, Chen G. A circRNA-miRNA-mRNA network identification for exploring underlying pathogenesis and therapy strategy of hepatocellular carcinoma. *J Transl Med*. 2018 Aug 9;16(1):220.
16. Sun Q, Liu Z, Xu X, Yang Y, Han X, Wang C, Song F, Mou Y, Li Y, Song X. Identification of a circRNA/miRNA/mRNA ceRNA Network as a Cell Cycle-Related Regulator for Chronic Sinusitis with Nasal Polyps. *J Inflamm Res*. 2022 Apr;Volume 15:2601–15.
17. Bai S, Wu Y, Yan Y, Shao S, Zhang J, Liu J, Hui B, Liu R, Ma H, Zhang X, Ren J. Construct a circRNA/miRNA/mRNA regulatory network to explore potential pathogenesis and therapy options of clear cell renal cell carcinoma. *Sci Rep*. 2020 Aug 12;10(1):13659.
18. Ma Y, Zou H. Identification of the circRNA-miRNA-mRNA Prognostic Regulatory Network in Lung Adenocarcinoma. *Genes*. 2022 May;13(5):885.
19. Cardenas J, Balaji U, Gu J. Cerina: systematic circRNA functional annotation based on integrative analysis of ceRNA interactions. *Sci Rep*. 2020 Dec 17;10:22165.
20. Chen Y, Yao L, Tang Y, Zhong JH, Wan J, Chang J, Cui S, Luo Y, Cai X, Li W, Chen Q, Huang HY, Wang Z, Chen W, Chang TH, Wei F, Lee TY, Huang HD. CircNet 2.0: an updated database for exploring circular RNA regulatory networks in cancers. *Nucleic Acids Res*. 2022 Jan 7;50(D1):D93–101.
21. Chen Z, Ren R, Wan D, Wang Y, Xue X, Jiang M, Shen J, Han Y, Liu F, Shi J, Kuang Y, Li W, Zhi Q. Hsa_circ_101555 functions as a competing endogenous RNA of miR-597-5p to promote colorectal cancer progression. *Oncogene*. 2019 Aug;38(32):6017–34.
22. Ding B, Yao M, Fan W, Lou W. Whole-transcriptome analysis reveals a potential hsa_circ_0001955/hsa_circ_0000977-mediated miRNA-mRNA regulatory sub-network in colorectal cancer. *Aging*. 2020 Mar 28;12(6):5259–79.

23. Quemener AM, Centomo ML, Sax SL, Panella R. Small Drugs, Huge Impact: The Extraordinary Impact of Antisense Oligonucleotides in Research and Drug Development. *Molecules*. 2022 Jan 15;27(2):536.
24. Hwang CL, Lai YJ, Liu TY. A new approach for multiple objective decision making. *Comput Oper Res*. 1993 Oct 1;20(8):889–99.
25. Barretina J, Caponigro G, Stransky N, Venkatesan K, Margolin AA, Kim S, Wilson CJ, Lehár J, Kryukov GV, Sonkin D, Reddy A, Liu M, Murray L, Berger MF, Monahan JE, Morais P, Meltzer J, Korejwa A, Jané-Valbuena J, Mapa FA, Thibault J, Bric-Furlong E, Raman P, Shipway A, Engels IH, Cheng J, Yu GK, Yu J, Aspesi P, de Silva M, Jagtap K, Jones MD, Wang L, Hatton C, Palesscandolo E, Gupta S, Mahan S, Sougnez C, Onofrio RC, Liefeld T, MacConaill L, Winckler W, Reich M, Li N, Mesirov JP, Gabriel SB, Getz G, Ardlie K, Chan V, Myer VE, Weber BL, Porter J, Warmuth M, Finan P, Harris JL, Meyerson M, Golub TR, Morrissey MP, Sellers WR, Schlegel R, Garraway LA. The Cancer Cell Line Encyclopedia enables predictive modelling of anticancer drug sensitivity. *Nature*. 2012 Mar;483(7391):603–7.
26. Glazár P, Papavasileiou P, Rajewsky N. circBase: a database for circular RNAs. *RNA* [Internet]. 2014 Sep 18 [cited 2021 Apr 19]; Available from: <http://rnajournal.cshlp.org/content/early/2014/09/18/rna.043687.113>
27. McGeary SE, Lin KS, Shi CY, Pham TM, Bisaria N, Kelley GM, Bartel DP. The biochemical basis of microRNA targeting efficacy. *Science*. 2019 Dec 20;366(6472):eaav1741.
28. Li JH, Liu S, Zhou H, Qu LH, Yang JH. starBase v2.0: decoding miRNA-ceRNA, miRNA-ncRNA and protein-RNA interaction networks from large-scale CLIP-Seq data. *Nucleic Acids Res*. 2014 Jan;42(Database issue):D92–97.
29. Garcia DM, Baek D, Shin C, Bell GW, Grimson A, Bartel DP. Weak seed-pairing stability and high target-site abundance decrease the proficiency of lsy-6 and other microRNAs. *Nat Struct Mol Biol*. 2011 Oct;18(10):1139–46.
30. Agarwal V, Bell GW, Nam JW, Bartel DP. Predicting effective microRNA target sites in mammalian mRNAs. *eLife*. 2015 Aug 12;4:e05005.
31. Soutschek M, Gross F, Schratt G, Germain PL. scanMiR: a biochemically based toolkit for versatile and efficient microRNA target prediction. *Bioinformatics*. 2022 May 1;38(9):2466–73.
32. Korotkevich G, Sukhov V, Budin N, Shpak B, Artyomov MN, Sergushichev A. Fast gene set enrichment analysis [Internet]. *bioRxiv*; 2021 [cited 2023 Mar 17]. p. 060012. Available from: <https://www.biorxiv.org/content/10.1101/060012v3>
33. Chen EY, Tan CM, Kou Y, Duan Q, Wang Z, Meirelles GV, Clark NR, Ma'ayan A. Enrichr: interactive and collaborative HTML5 gene list enrichment analysis tool. *BMC Bioinformatics*. 2013 Apr 15;14:128.
34. Ritchie ME, Phipson B, Wu D, Hu Y, Law CW, Shi W, Smyth GK. limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Res*. 2015 Apr 20;43(7):e47–e47.
35. Duan X, Yu X, Li Z. Circular RNA hsa_circ_0001658 regulates apoptosis and autophagy in gastric cancer through microRNA-182/Ras-related protein Rab-10 signaling axis. *Bioengineered*. 2022 Feb;13(2):2387–97.
36. Lu Y, Wang XM, Li ZS, Wu AJ, Cheng WX. Hsa_circ_0001658 accelerates the progression of colorectal cancer through miR-590-5p/METTL3 regulatory axis. *World J Gastrointest Oncol*. 2023 Jan 15;15(1):76–89.
37. Jiang F, Shen XB. miRNA and mRNA expression profiles in gastric cancer patients and the relationship with circRNA. *Neoplasma*. 2019 Nov;66(6):879–86.
38. Zhou D, Dong L, Yang L, Ma Q, Liu F, Li Y, Xiong S. Identification and analysis of circRNA–miRNA–mRNA regulatory network in hepatocellular carcinoma. *IET Syst Biol*. 2020 Nov 25;14(6):391–8.
39. Wang Z, Pei H, Liang H, Zhang Q, Wei L, Shi D, Chen Y, Zhang J. Construction and Analysis of a circRNA-Mediated ceRNA Network in Lung Adenocarcinoma. *OncoTargets Ther*. 2021 Jun 8;14:3659–69.
40. Liang L, Zhang L, Zhang J, Bai S, Fu H. Identification of circRNA-miRNA-mRNA Networks for Exploring the Fundamental Mechanism in Lung Adenocarcinoma. *OncoTargets Ther*. 2020 Apr 8;13:2945–55.
41. Liu Y, Wang X, Bi L, Huo H, Yan S, Cui Y, Cui Y, Gu R, Jia D, Zhang S, Cai L, Li X, Xing Y. Identification of Differentially Expressed Circular RNAs as miRNA Sponges in Lung Adenocarcinoma. *J Oncol*. 2021 Sep 10;2021:5193913.
42. He S, Guo Z, Kang Q, Wang X, Han X. Circular RNA hsa_circ_0000517 modulates hepatocellular carcinoma advancement via the miR-326/SMAD6 axis. *Cancer Cell Int*. 2020 Aug 3;20:360.
43. Wang X, Wang X, Li W, Zhang Q, Chen J, Chen T. Up-Regulation of hsa_circ_0000517 Predicts Adverse Prognosis of Hepatocellular Carcinoma. *Front Oncol*. 2019 Oct 22;9:1105.
44. Xu Y, Yu J, Huang Z, Fu B, Tao Y, Qi X, Mou Y, Hu Y, Wang Y, Cao Y, Jiang D, Xie J, Xu Y, Zhao J, Xiong W. Circular RNA hsa_circ_0000326 acts as a miR-338-3p sponge to facilitate lung adenocarcinoma progression. *J Exp Clin Cancer Res CR*. 2020 Apr 5;39:57.

Session 5: Algorithms and data structures for sequences / Knowledge representation

HairSplitter: separating strains in metagenome assemblies with long reads

Roland FAURE^{1,2}, Jean-François FLOT² and Dominique LAVENIER¹

¹ Univ. Rennes, INRIA RBA, CNRS UMR 6074, Rennes, France

² Service Evolution Biologique et Ecologie, Université libre de Bruxelles (ULB), 1050 Brussels, Belgium

Corresponding author: roland.faure@irisa.fr

Abstract *Long read assemblers struggle to distinguish closely related strains of the same species and collapse them into a single sequence. This is very limiting when analysing a metagenome, as different strains can have important functional differences. We present the first version of a new software called HairSplitter, which recovers the strains from a strain-oblivious assembly and long reads. The originality of the method lies in a custom variant calling step that allows HairSplitter to work with erroneous reads and to separate an unknown number of haplotypes. On simulated datasets, we show that HairSplitter significantly outperforms the state of the art when dealing with metagenomes containing many strains of the same species.*

Keywords Metagenomics, Haplotyping, Genome assembly, Strain separation

1 Introduction

A powerful tool for understanding complex microbial communities is de novo metagenome assembly. Current methods can reconstruct the genomes of sufficiently abundant species, but struggle to differentiate strains within a species, even if they are abundant. While strains of the same species are very similar at the genomic level, the small differences can lead to very significant phenotypic and functional changes. The most famous example of such intra-specific diversity is probably *Escherichia coli* [1], some strains of which can be highly pathogenic while sharing an average nucleotide identity of more than 98.5% with commensal strains [2].

Assemblers are designed to correct for sequencing errors by ignoring bases that occur at low frequencies. As a side effect, they generally discard haplotype differences and collapse the different haplotypes into a single sequence. An additional difficulty in the metagenomic context is that the number of haplotypes is a priori unknown and that haplotypes generally have different frequencies in a sample.

Specific software has been developed to overcome these difficulties. For example, two such software based on short reads are STRONG [3] and strainXpress [4]. However, more and more samples are sequenced using only long reads, as their cost has dropped recently and they allow for more contiguous assemblies.

Using error-prone long reads, assemblers such as metaFlye [5] or Canu [6] attempt to assemble strains separately. However, the authors of [7] showed that these assemblers still struggle to recover multiple strains and proposed a new pipeline, called *Strainberry*, which takes an assembly and the long reads as input and recovers the collapsed strains. Strainberry improves significantly the assembly of samples containing 2 or 3 strains of the same species, but is limited when the number of species increases.

We present *HairSplitter*, a new pipeline for recovering the strains lost when assembling exclusively from (error-prone) long reads. HairSplitter does not make any assumption on the number of strains that should be found in the metagenome. It contains an original procedure that combines a custom variant calling method with a new phasing algorithm. We have extensively tested HairSplitter on simulated Nanopore data, replicating the protocol proposed in [7], and show that HairSplitter significantly improves the completeness of assemblies of metagenomes composed of many strains compared to the state of the art. HairSplitter still needs to be tested on real datasets.

2 Description of the pipeline

The HairSplitter pipeline is composed of four main steps: 1) alignment of the reads on the contigs, 2) separation of the reads in their haplotype of origin, 3) generation of the new contigs and 4) strain-aware contig scaffolding. Steps 2 is done by an original software, while step 1 is performed by minimap2 [8], step 3 by Racon [9] and step 4 by GraphUnzip [10]. The pipeline is illustrated Figure 2

Step 1: Aligning the reads on the contigs

HairSplitter starts by generating base-to-base alignments of the sequencing reads on the assembly with minimap2. For each contig, a multiple sequence alignment is generated using the alignment of the reads to the reference. This is the simplest way to build a multiple sequence alignment, but it creates alignment artifacts, especially at positions where the contig contains errors.

Step 2: Splitting a group of reads in one or more haplotypes

The originality of HairSplitter lies in the second and most crucial step, where reads that align on a contig are separated by haplotype of origin.

This operation is performed locally on window of the contig of size w . The phasing is performed locally to avoid clustering reads that do not overlap. Indeed, clustering together reads that do not overlap could lead to the HairSplitter algorithm clustering very different reads together, as shown in Figure 1. w should thus be chosen to be significantly shorter than the reads.

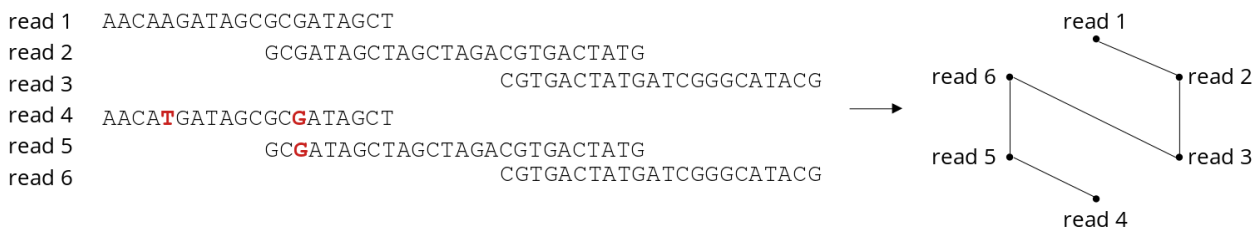


Fig. 1. A read graph is built as follows: each read is a vertex ; reads are connected to the reads they overlaps with 100% identity. Even though read1 and read4 are very different, they are transitively linked through read3 and read6.

A difficulty of the process comes from the error rate of the reads, which can be much higher than the divergence between the haplotypes. Another difficulty is the possibly high number of haplotypes which can be unevenly covered by the sequencing. As this step represents the core of HairSplitter, it is described in detail in section 3.

Step 3: Generating new contigs

The reads on a given window are separated into n groups, the contig sequence in the window is polished n times by Racon using the different groups, yielding n different versions of the window, which we call *subcontigs*. The subcontigs are laid out as an assembly graph, based on the original assembly graph.

Step 4: Strain-aware scaffolding

Due to local homozygosity, some subcontigs will contain multiple haplotypes. These subcontigs limit the contiguity of the graph and must be duplicated to be present once for each haplotype. To do this, the paths of the sequencing reads on the subcontig graph are inventoried. Once all the paths are inventoried, GraphUnzip [10] untangles the graph. From the paths of the read in the graph, it deduces which contig to duplicate and which contig to link to improve the contiguity and completeness of the assembly.

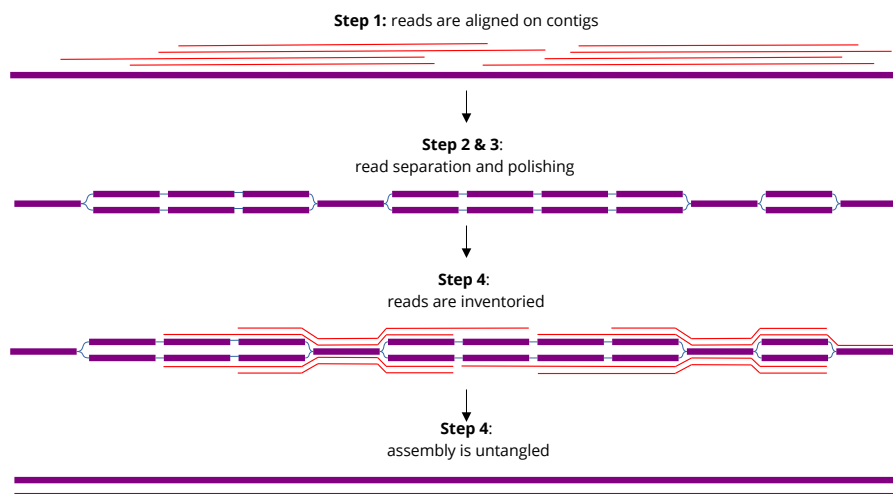


Fig. 2. The HairSplitter pipeline. The purple rectangles represent contigs or subcontigs. The red lines are reads used to build the subcontigs. Only a very small subset of reads is shown here to keep the visualisation readable.

3 Splitting a group of reads in one or more haplotypes

3.1 Detecting variants

To separate the set of reads that align on a contig in several haplotypes, rudimentary variant calling is performed. In this context, we define variants as positions where the different haplotypes are not identical. Once the variants are clearly identified, the reads can be split into the different haplotypes based on these positions, as can be seen in Figure 3. However, due to errors in the reads and in the reference, differences between read and reference do not systematically highlight a variant, as can be seen in Figure 4.

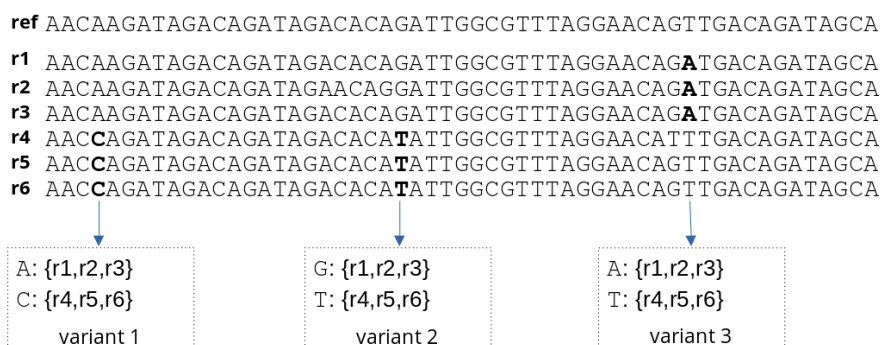


Fig. 3. In this error-free alignment, the variants clearly separate the reads in two haplotypes, one containing r1, r2 and r3 and the other containing r4, r5 and r6.

3.2 Dealing with errors

To distinguish variants from errors, all positions of the alignment are iteratively inspected. For each position, the program inventories the triplet of bases centered there in the reads. A base triplet is defined as the base at the given position flanked by the nucleotide on the left and the nucleotide on the right in the given read. At conserved regions, which make up the majority of the contig, the reference triplet will be the most common, while some residual triplets may be due to sequencing errors or alignment artifacts. At positions of true genomic variants, two triplets (corresponding to the two variants) will be common, with some residual triplets due to sequencing errors or alignment artifacts. *Suspect positions* are defined as positions where the second most abundant triplet is significantly (by a factor s) more abundant than the third most abundant triplet. All variant should generate at least one suspicious positions, therefore only suspect positions are considered for phasing. This will discard most positions where there are only a few random errors. Some positions where there are no true

variants will also be flagged as suspicious, typically positions with alignment artifacts. The selection of suspicious position is illustrated Figure 4.

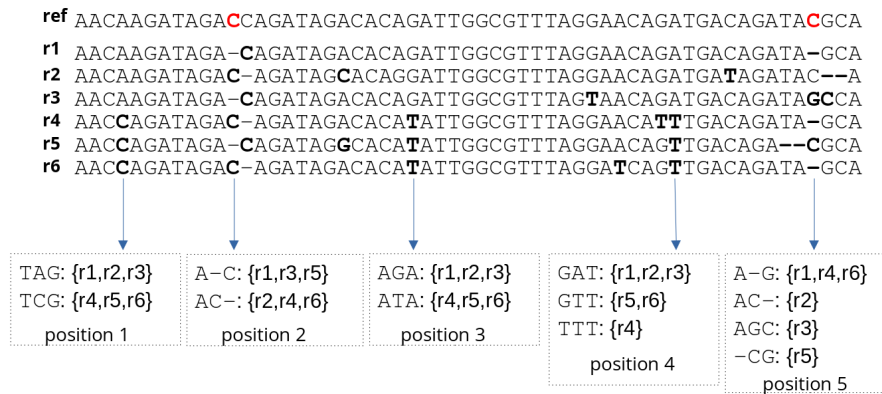


Fig. 4. Selection of suspicious positions. The red bases in the reference represent errors (unknown to HairSplitter). Note that position 2 will be flagged as suspicious because of an alignment artifact. Position 5 will not be considered suspicious because even though there are many different triplets at this position, there is no clear alternative to the triplet ‘A-G’.

To distinguish variants from alignment artifacts, HairSplitter implements a method based on the intuition that alignment artifacts are generally randomly distributed on the reads, while genomic variants are not. On the one hand, reads carrying sequencing and alignment errors are not correlated between two positions. On the other hand, reads carrying variants should be strongly correlated between two variants.

Suspicious positions are clustered hierarchically. A cluster is represented by its consensus. Two clusters are merged if more than 90% of each part of each consensus consists of reads from a part of the other consensus. At the end of the process, consensus consisting of more than p positions are considered solid. We call them the *solid bipartitions* of the reads.

For each suspect position, its correlation with all solid bipartitions is computed by a one-degree-of-freedom chi-square test of independence. If the result is greater than five (strong correlation), the position is marked as *interesting*. The positions that do not correlate well with any solid bipartitions are considered untrustworthy and are discarded. The set of interesting positions corresponds to the set of positions that will be considered as a bi-allelic variants by the algorithm.

This variant calling procedure is quite conservative and may miss existing variants. This is not a problem, as the goal of this step is to confidently identify a set of variants, not to call all variants exhaustively.

3.3 Phasing the reads using the variants

Once a set of variants has been largely extracted from the noise, reads are split into different haplotypes.

A graph is generated for each window of the contig. The reads that cover the window from end to end are the vertices of the graph. Pairwise distances between the reads are calculated as the number of divergent interesting positions divided by the total number of interesting positions overlapped by both reads. Each read is connected to its k nearest neighbors, as shown Figure 5.

The resulting graph forms clusters, grouping reads that share identical variants, corresponding to haplotypes. Empirically, the best algorithm to cluster this graph without knowing a priori the number of clusters seems to be the Chinese Whispers algorithm [11]. However, a limitation of this algorithm is that it is not deterministic, especially when the number of nodes in the graph is small. With low frequency, clusters will be merged or split. To avoid this, HairSplitter exploits the property that if the Chinese Whispers algorithm is initialized with an approximate solution, it will converge to the solution without splitting or merging clusters. The approximate solutions can be found at the interesting positions: each position contains a variant that separates two groups of reads with some

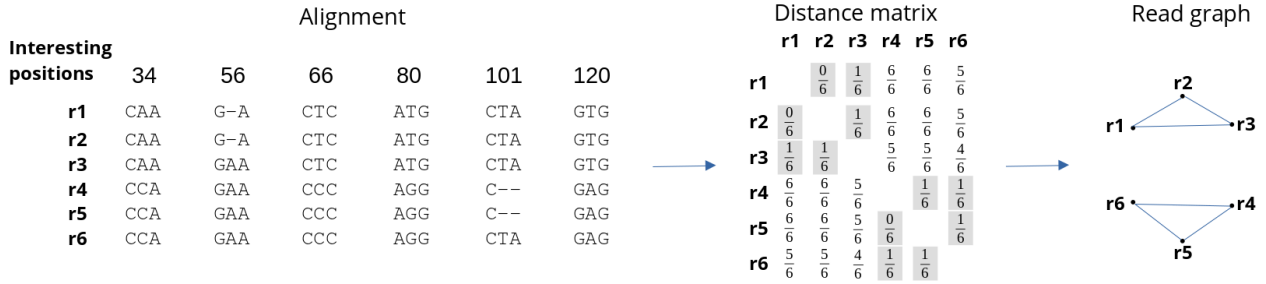


Fig. 5. Generating the read graph from the list of interesting positions with $k = 2$. Light gray squares highlight the k closest neighbors of each read in the distance matrix, which become ones in the adjacency matrix of the read graph.

errors. Running the Chinese whispers algorithm with one position as initialization will cluster the graph into two parts. Each interesting positions will yield a bipartition. All the bipartitions can then be aggregated into a single partition: two reads will be clustered together if and only if they are not separated in any bipartition. This process is illustrated Figure 6.

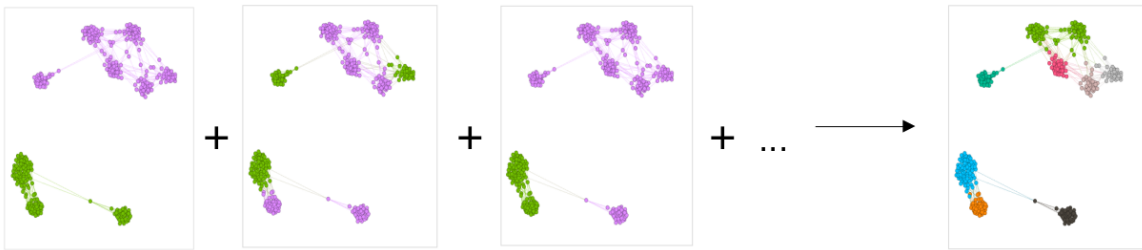


Fig. 6. The read graph is clustered using interesting positions as initialisation, resulting in a series of bipartitions. All these bipartitions can then be aggregated into the final partition.

Each part of the resulting partition corresponds to a haplotype.

4 Results

4.1 Protocol

The protocol is a replica of the protocol proposed in [7].

To systematically test the performance of HairSplitter, we composed a benchmark consisting of different strains of *Escherichia coli*. The reference genomes of the strains were obtained from NCBI. For each strain, we simulated “mediocre” Nanopore sequencing (around 7% error rate), using Badreads [12] with default settings, from the reference genomes.

For each experiment, the sequencing of different strains was concatenated to create a simulation of the sequencing of a metagenomic sample. The reads were then assembled using metaFlye with default settings. Missing strains were then recovered from the assembly using Strainberry and HairSplitter with default settings. For HairSplitter, the parameters are set to $s = 5$, $k = 5$, $p = 5$ and $w = 2000$, but the few tests we ran suggest that the algorithm is not very sensitive to these settings.

4.2 Evaluation metrics

Two metrics were retained to evaluate the quality of the recovered assembly with respect to the known solution genomes.

The first one is the proportion of the 21-mers found in the genomes that are not found in the assembly. This measures how well the different strains are covered. A large number of missing 21-mers indicates that some strains have not been well assembled.

The second metric is the proportion of 21-mers found in the assembly that are found in the genome. This evaluates the accuracy of the assembly. A low number of 21-mers found in the genomes indicates that the assembly contains many errors.

4.3 Influence of strain coverage, divergence and number of strains on strain separation

Divergence A factor that can influence the strain reconstruction is the degree of divergence between the strains. Seven datasets were created, composed of the simulated sequencing of the K12 strain mixed with the simulated sequencing of seven strains having varying degree of divergence with K12. The strains have been chosen to replicate exactly the experiment shown in [7].

Coverage Another crucial factor to reconstruct the strains is the depth at which each strain is covered. To evaluate how this affected the HairSplitter algorithm, tests were carried out on a mixture of the IAI1 and 12009 strains. In a first experiment, the two strains were sequenced at depths ranging from 5x to 50x. In a second experiment, the 12009 strain was sequenced systematically at 50x coverage, while the IAI1 strain was sequenced at depths ranging from 5x to 50x.

Number of strains The authors of Strainberry pointed to the number of strains as a limiting factor in strain reconstruction, with the completeness of the reconstruction decreasing significantly when more than 3 strains were sequenced [7]. Mixtures were made with different numbers of strains. The strains used for the mixtures were 12009, IAI1, F11, S88, Sakai, SE15, *Shigella flexneri*, UMN026, HS and K12. The strains were chosen to cover a wide range of the *Escherichia coli* phylogenetic tree, with some very close strains (such as F11 and S88) and others much more distance strains (such as SE15 and K12).

The results Figure 7 show that HairSplitter and Strainberry recover a very similar amount of k-mers on mixtures of 2 strains. More specifically, both software perform well on pairs of strains with more than 0.3% divergence, as defined by the ANI [13] (Figures 7b), and when the coverage is 20x or more (Figures 7d, 7f). This confirms the results presented in [7]. However, while HairSplitter recovers a similar amount of missing 21-mers compared to Strainberry, it tends to generate much fewer erroneous 21-mers (Figure 7a, 7c and 7e).

The most spectacular result is that even when the mixture contained six or more strains, HairSplitter was able to recover most of the missing 21-mers, producing assemblies with significantly fewer missing 21-mers than metaFlye and even Strainberry assemblies (Figure 7h). For example, in the metaFlye assembly of the mixture of 10 strains, 46% of the 21-mers of the solution were missing. This dropped to 39% when using Strainberry and 16% when using HairSplitter.

4.4 Performance

In all these tests, HairSplitter finished in less than 30 minutes using four threads and less than 5G of RAM. This is similar to Strainberry and small compared to the assembly time, which took at least 5 times longer.

5 Discussion

In this work, we introduced HairSplitter, a new pipeline for performing strain separation on assemblies using only long reads. HairSplitter shows a significant improvement over state-of-the-art methods when the number of strains is high. One of the reasons for this is that, unlike Strainberry, HairSplitter can separate a contig into a variable number of strains. To confirm these results, HairSplitter needs to be benchmarked on real sequencing datasets.

The data we simulated was of low quality to test HairSplitter in the hardest possible case. We expect HairSplitter to perform even better as the quality of the sequencing improves, but this remains to be tested. We also want to see if HairSplitter can improve on the de novo assembly performed by hifiasm in the case of HiFi sequencing.

Another potential application of HairSplitter that deserves investigation is the phasing of polyploid species. Powerful software, such as WhatsHap [14], already exists when the number of haplotypes is known a priori and can be used for polyploid assembly. However, knowing the number of haplotypes in each contig is not necessarily an easy task and using an agnostic approach such as HairSplitter could improve the results.

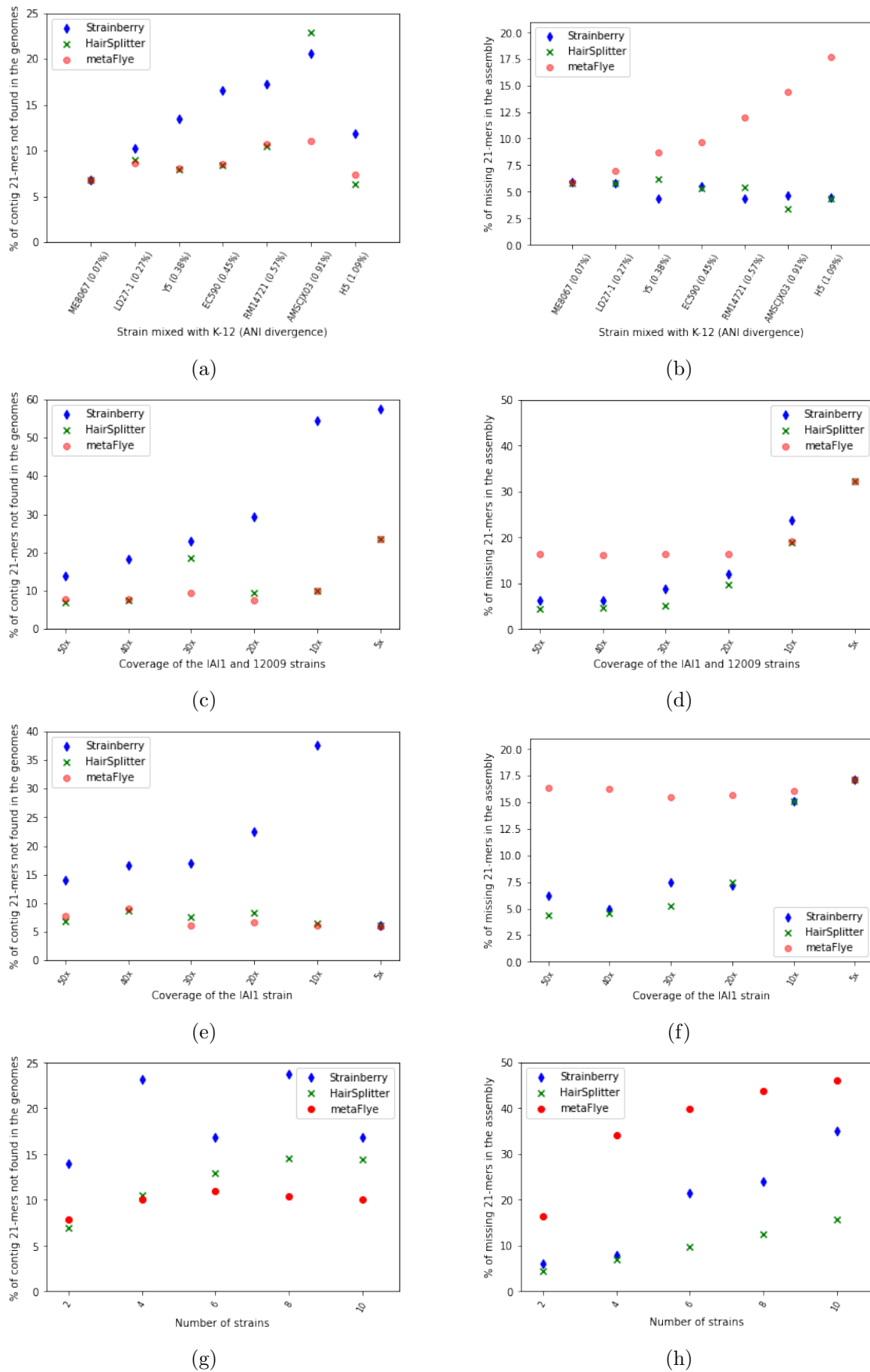


Fig. 7. Evaluation of assemblies obtained using HairSplitter or Strainberry on the metaFlye assembly in different mixes of strains. a and b: mix of K12 strain and another strain at 50x coverage. c and d: mix of the IAI1 and 12009 strains at varying coverage. e and f: mix of the IAI1 and 12009 strains, with 12009 at 50x coverage and IAI1 at varying coverage. g and h: mix of varying number of strains at 50x coverage.

Currently, HairSplitter has two shortcomings that we are trying to fix as a priority. First, it needs at least 20x coverage to distinguish a strain. This is quite high and will undoubtedly be limiting in most real-world applications where rare strains are common. Improving the quality of the data may solve this problem. The second shortcoming is that the contig scaffolding step is still not fully satisfactory and the contiguities of the assemblies obtained is lower than those obtained using Strainberry. This is due to the fact that we are using GraphUnzip slightly outside the use case for which it was designed. We will adapt GraphUnzip to this specific task.

Acknowledgements

We thank the Genouest facility (genouest.org) for providing us the hardware, software and support necessary to run our computations. The software Tablet [15] and Bandage [16] were used to visualize data while developing HairSplitter. For the purpose of Open Access, a CC-BY public copyright licence has been applied by the authors to the present document and will be applied to all subsequent versions up to the Author Accepted Manuscript arising from this submission

References

- [1] Olivier Tenaillon, David Skurnik, Bertrand Picard, and Erick Denamur. The population genetics of commensal *Escherichia coli*. *Nature Reviews Microbiology*, 8(3):207–217, March 2010.
- [2] S Hudault. *Escherichia coli* strains colonising the gastrointestinal tract protect germfree mice against *Salmonella typhimurium* infection. *Gut*, 49(1):47–55, July 2001.
- [3] Christopher Quince, Sergey Nurk, Sebastien Raguideau, Robert James, Orkun S. Soyer, J. Kimberly Summers, Antoine Limasset, A. Murat Eren, Rayan Chikhi, and Aaron E. Darling. Metagenomics Strain Resolution on Assembly Graphs. preprint, Bioinformatics, September 2020.
- [4] Xiongbin Kang, Xiao Luo, and Alexander Schönhuth. StrainXpress: strain aware metagenome assembly from short reads. *Nucleic Acids Research*, 50(17):e101–e101, September 2022.
- [5] Mikhail Kolmogorov, Derek M. Bickhart, Bahar Behsaz, Alexey Gurevich, Mikhail Rayko, Sung Bong Shin, Kristen Kuhn, Jeffrey Yuan, Evgeny Pevnikov, Timothy P. L. Smith, and Pavel A. Pevzner. metaFlye: scalable long-read metagenome assembly using repeat graphs. *Nature Methods*, 17(11):1103–1110, November 2020.
- [6] Sergey Koren, Brian P. Walenz, Konstantin Berlin, Jason R. Miller, Nicholas H. Bergman, and Adam M. Phillippy. Canu: scalable and accurate long-read assembly via adaptive k -mer weighting and repeat separation. *Genome Research*, 27(5):722–736, May 2017.
- [7] Riccardo Vicedomini, Christopher Quince, Aaron E. Darling, and Rayan Chikhi. Strainberry: automated strain separation in low-complexity metagenomes using long reads. *Nature Communications*, 12(1):4485, July 2021.
- [8] Heng Li. Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics*, 34(18):3094–3100, September 2018.
- [9] Li Fang and Kai Wang. Polishing high-quality genome assemblies. *Nature Methods*, 19(6):649–650, June 2022.
- [10] Roland Faure, Nadège Guiglielmoni, and Jean-François Flot. GraphUnzip: unzipping assembly graphs with long reads and Hi-C. preprint, Bioinformatics, February 2021.
- [11] Chris Biemann. Chinese whispers: An efficient graph clustering algorithm and its application to natural language processing problems. *Proceedings of TextGraphs*, pages 73–80, 07 2006.
- [12] Ryan Wick. Badread: simulation of error-prone long reads. *Journal of Open Source Software*, 4(36):1316, April 2019.
- [13] Johan Goris, Konstantinos T. Konstantinidis, Joel A. Klappenbach, Tom Coenye, Peter Vandamme, and James M. Tiedje. DNA–DNA hybridization values and their relationship to whole-genome sequence similarities. *International Journal of Systematic and Evolutionary Microbiology*, 57(1):81–91, January 2007.
- [14] Sven D. Schrunner, Rebecca Serra Mari, Jana Ebler, Mikko Rautiainen, Lancelot Seillier, Julia J. Reimer, Björn Usadel, Tobias Marschall, and Gunnar W. Klau. Haplotype threading: accurate polyploid phasing from long reads. *Genome Biology*, 21(1):252, September 2020.
- [15] I. Milne, G. Stephen, M. Bayer, P. J. A. Cock, L. Pritchard, L. Cardle, P. D. Shaw, and D. Marshall. Using Tablet for visual exploration of second-generation sequencing data. *Briefings in Bioinformatics*, 14(2):193–202, March 2013.
- [16] Ryan R. Wick, Mark B. Schultz, Justin Zobel, and Kathryn E. Holt. Bandage: interactive visualization of *de novo* genome assemblies. *Bioinformatics*, 31(20):3350–3352, October 2015.

Opening the Black Box of Imputation Software to Study the Impact of Reference Panel Composition on Performance

Thibault DEKEYSER^{1,2}, Emmanuelle GÉNIN² and Anthony HERZIG²

¹ Inserm, Université de Brest, EFS, UMR 1078, GGB, F-29200 Brest, France

² CHRU Brest, F-29200 Brest, France

Corresponding Author: anthony.herzig@inserm.fr

Dekeyser T, Génin E, Herzig AF. Opening the Black Box of Imputation Software to Study the Impact of Reference Panel Composition on Performance. *Genes*. 2023; 14(2):410. <https://doi.org/10.3390/genes14020410>

Genotype imputation is widely used to enrich genetic datasets. The operation relies on panels of known reference haplotypes, typically with whole-genome sequencing data. How to choose a reference panel has been widely studied and it is essential to have a panel that is well matched to the individuals who require missing genotype imputation. However, it is broadly accepted that such an imputation panel will have an enhanced performance with the inclusion of diversity (haplotypes from many different populations). We investigate this observation by examining, in fine detail, exactly which reference haplotypes are contributing at different regions of the genome. This is achieved using a novel method of inserting synthetic genetic variation into the reference panel in order to track the performance of leading imputation algorithms.

We took data from the 1000 Genomes Project as our sandbox. We decided to separate three populations (ACB, ASW, and MXL) and impute the 221 individuals from these populations (our target group) with a reference panel formed by the remaining 23 populations. In order to track which reference haplotypes were being called on to impute the individuals of our target group across the genome, we injected completed synthetic variants into the imputation reference panel. These variants would serve as indicators for each of the five continental reference groups; a synthetic variant tagging the EUR population would be 0 (the ‘reference’ allele) for all non-EUR haplotypes and 1 (the ‘alternative’) for all EUR haplotypes. We refer to each batch of five synthetic variants as an ‘imputation barcode’.

Having added 32,279 imputation barcodes, imputation was completed using IMPUTE5 . We could then calculate the cumulative contributions of different reference groups to the imputation. This was compared to similar estimations using the chromo-painting functionality of pbwt , supervised ADMIXTURE , and SOURCEFIND . Based on these results, we performed a large number of further imputations using restricted imputation panels. This led to a proposed optimisation strategy where the imputation panel would be restricted differently in each genomic region. This allowed us to demonstrate clearly the role of diversity in an imputation reference panel, explaining previous results in the literature as well as providing potential avenues for improving existing genotype imputation algorithms. We show that while diversity may globally improve imputation accuracy, there can be occasions where incorrect genotypes are imputed following the inclusion of more diverse haplotypes in the reference panel.

References

Session 6: Workflows, reproducibility and open science

SnakeMAGs: a simple, efficient, flexible and scalable workflow to reconstruct prokaryotic genomes from metagenomes

Nachida Tadrent¹, Franck Dedeine¹, and Vincent Herve^{*†1,2}

¹Institut de recherche sur la biologie de l'insecte UMR7261 – Université de Tours, Centre National de la Recherche Scientifique, Centre National de la Recherche Scientifique : UMR7261 – France

²Paris-Saclay Food and Bioproduct Engineering – AgroParisTech, Université Paris-Saclay, Institut National de Recherche pour l'Agriculture, l'Alimentation et l'Environnement – France

Résumé

Over the last decade, microbial ecology has undergone a transition from gene-centric to genome-centric analyses. Indeed, the advent of metagenomics combined with binning methods, single-cell genome sequencing as well as high-throughput cultivation methods have contributed to the continuing and exponential increase of available prokaryotic genomes, which in turn has favoured the exploration of microbial metabolisms. In the case of metagenomics, data processing, from raw reads to genome reconstruction, involves various steps and software which can represent a major technical obstacle.

To overcome this challenge, we developed *SnakeMAGs*, a simple workflow that can process Illumina data, from raw reads to metagenome-assembled genomes (MAGs) classification and relative abundance estimate. It integrates state-of-the-art bioinformatic tools to sequentially perform: quality control of the reads (illumina-utils, Trimmomatic), host sequence removal (optional step, using Bowtie2), assembly (MEGAHIT), binning (MetaBAT2), quality filtering of the bins (CheckM, GUNC), classification of the MAGs (GTDB-Tk) and estimate of their relative abundance (CoverM). Developed with the popular Snakemake workflow management system (1), it can be deployed on various architectures, from single to multicore and from workstation to computer clusters and grids. It is also flexible since users can easily change parameters and/or add new rules.

Using termite gut metagenomic datasets, we showed that *SnakeMAGs* is slower but allowed the recovery of more MAGs encompassing more diverse phyla compared to another similar workflow named ATLAS (2). Importantly, these additional MAGs showed no significant difference compared to the other ones in terms of completeness, contamination, genome size nor relative abundance. Overall, our workflow should make the reconstruction of MAGs more accessible to microbiologists.

SnakeMAGs is currently used in several large-scale metagenomic projects from different institutions. To offer more freedom and flexibility to the users, future releases of the workflow will include a larger choice of tools to perform the same task (e.g. different trimming, assembly or binning software). SnakeMAGs as well as test files and an extended tutorial are available at: <https://github.com/Nachida08/SnakeMAGs>.

*Intervenant

†Auteur correspondant: vincent.herve@inrae.fr

Mots-Clés: Snakemake, metagenomics, microbiology, genomics, microbial ecology

BEAURIS: an automated, modular, and FAIR system for large-scale genome data management

Matéo BOUDET^{1,2}, Loraine BRILLET-GUÉGUEN^{3,4}, Arthur LE BARS^{3,5}, Karine MASSAU^{3,6}, Laura LEROI⁷, Alexandre CORMIER⁷, Patrick DURAND⁷, Erwan CORRE³, Anthony BRETAUDEAU^{1,2}

¹ IGEPP, Agrocampus Ouest, INRAE, Université de Rennes 1, 35653 Le Rheu, France

² INRIA, IRISA, GenOuest Core Facility, Campus de Beaulieu, 35000 Rennes, France

³ CNRS, Sorbonne Université, FR2424, ABiMS-IFB, Station Biologique, 29680, Roscoff, France

⁴ Sorbonne Université, CNRS, Integrative Biology of Marine Models (LBI2M), Station Biologique de Roscoff (SBR), 29680, Roscoff, France

⁵ CNRS, Institut Français de Bioinformatique, IFB-core, UAR 3601, Évry, France

⁶ Université de Rouen Normandie, 76130, Mont-Saint-Aignan, France

⁷ Ifremer, Service de Bioinformatique (SeBiMER), ZI de la Pointe du Diable, 29280 Plouzané, France

Corresponding Author: mateo.boudet@inrae.fr

Keywords: Reproducibility, data integration, automation, genome, annotation

Introduction

The rapid democratization of genome sequencing technologies has led to a significant growth in genomic data available for non-model species in recent years. However, the massive scale and diversity of the produced data pose significant challenges for scientists to meaningfully explore and extract information. To address these challenges, there is a need for robust information systems that can efficiently integrate, analyze, and visualize genomic data from multiple heterogeneous sources. These systems must be capable of handling the increasing complexity and interconnectivity of genomic data, managing data security, and providing intuitive interfaces for users to explore and interact with the content.

Although international databanks exist for genomic data (e.g. NCBI [1], EBI [2], Ensembl [3]), they may not fully meet the needs of all users and communities, in terms of integration, analysis and visualization capabilities. Additionally, these databanks typically have specific data formats and access policies, which may not be ideal for all research. In response, more specialized databanks have emerged to address the specific needs of communities. They are often based on open-source software (e.g. GMOD tool suite [4]), and provide tailored interfaces, data models, and annotation resources. However, the increasing pace of data release poses a significant challenge: as the volume and complexity of data increase, so do the human and technical resources required for their curation and integration, which can be difficult for smaller organizations to sustain.

For three years the Breton teams of the BioGenouest bioinformatics axis have been working together implementing portals for the provision, visualization and processing of genome data [5]. In order to federate the work carried out around common developments, a working group was formed to set up a FAIR [6] compatible, automated system for data curation and integration. BEAURIS is the result of this work.

Results

In this talk, we will present BEAURIS [7], a fully automated system for integration, visualization and exploration of genomic data. The target audience of BEAURIS are bioinformaticians willing to publish genomic data within web interfaces.

BEAURIS is the combination of GitLab CI/CD [8] as an automated pipeline engine, and a custom-made, modular Python library for the management and processing of the data. The pipeline itself works in four steps: i) data validation and correction (e.g. reformatting of common GFF formatting issues), ii) data

derivation (e.g. format conversion, metrics computing, functional annotation with ORSON [9]), iii) web interface deployment, and iv) data safeguarding, with each step being composed of several configurable jobs. As CI/CD computing resources are limited, long running steps can be executed on external computing platforms (Galaxy, HPC clusters with DRMAA, using or not workflow systems like NextFlow) thanks to the usage of specifically configured GitLab Runners [10].

Adding new genomic data into BEAURIS consists in writing yaml files, conforming to a well-defined schema. These files contain paths to the data itself, various metadata, and optional parameters to allow full customization of the pipeline. For instance, the user may select which web interfaces to deploy from a selection (including JBrowse [11], Apollo [12], Blast [13,14], GeneNoteBook [15], and a download page), restrict access to the data to a specific group of user, or modify the amount of computational resources used by the pipeline. The jobs launched by BEAURIS will vary depending on the data provided, (eg: assembly, annotation, and/or track files). A versioning system allows to add multiple versions of each dataset, and to keep track of all the history of this data.

On submission of the yaml files (through a merge request), a staging pipeline is launched, following the four steps described previously. On pipeline completion, the user can access the processed data and the web interfaces in a staging environment, accessible only to admin users, allowing to preview the final result, and if needed amend the submitted yaml files. Any change in the yaml files will be detected, and will re-trigger the required jobs depending on the changes.

At this point, the merge request may be accepted by the administrator, and will trigger the production pipeline. This pipeline will lock both the initial and processed data, labeling it with metadata, and store it in a dedicated safe folder, with restricted permissions, to ensure it cannot be tampered with in the long term. The web interfaces will also be deployed in the production environment at this step, with public access to anyone, or restricted to specific audiences.

At any point after the merge, the user may submit another merge request to amend the yaml file, by adding for instance a new assembly or annotation. The same pipeline will then be triggered, and only the jobs linked to the modified or added data will run.

Currently, BEAURIS has already been used to integrate data from 9 distinct organisms (genome sequences, structural and functional annotations, RNASeq data), on two bioinformatics platforms (GenOuest, ABiMS).

Conclusion

BEAURIS has been designed with a strong will to respect the principles of FAIR data and open science. Indeed, the code, and the chosen architecture is by nature embracing the open science concept (free access to the code and the infrastructure, open license). The use of structured yaml files, and the long term safe storage of raw and processed data, make BEAURIS a very strong basis for publishing genomic data in reproducible and FAIR manner. The automatic submission of data to external data repositories (e.g. Zenodo, recherche.data.gouv.fr, NCBI, EBI) is already planned in the near future.

Technically, future versions of this system will allow to support a greater range of data types (e.g. genomic variants, orthology and synteny data, phenotypic data) and web interfaces (e.g. AskOmics, JBrowse2, synteny viewers). Thanks to the modularity of BEAURIS, each bioinformatics platform will be able to contribute their own components, and to use the ones they need for their specific audiences.

Finally, BEAURIS will be the basis for a 8-year major project starting this year. The ATLASea programme (PEPR) plans to sequence the genome of 4500 of the 12,000 marines eukaryotic species identified in the metropolitan EEZ and 1,000 species in four overseas territories. The BEAURIS project partners will all be involved in the targeted BYTE-SEA project to set up the IT infrastructure dedicated to the management and analysis of the programme's data. The aim will be to centralize all genomic data produced in a single portal and to integrate the genomes of related organisms sequenced by other consortia. The portal will provide tools for sequence searching, genome analysis, genome comparison and data visualization. In this context, all the developments carried out over the last few years within the framework of BEAURIS will make it possible to ensure the programmatic implementation of genome visualization spaces in line with the high throughput of raw data, estimated at 3 genomes per day. On the other hand, it will guarantee the interoperability and

security of the data, and facilitate their dissemination and use in accordance with FAIR and Open Science principles.

Code availability

Code available under MIT license on <https://gitlab.com/beaur1s/beauris>, contributions are welcome.

References

- [1] Sayers EW, Bolton EE, Brister JR, Canese K, Chan J, Comeau DC, et al. Database resources of the national center for biotechnology information. *Nucleic Acids Res* 2022;50:D20–6. <https://doi.org/10.1093/nar/gkab1112>.
- [2] Kanz C. The EMBL Nucleotide Sequence Database. *Nucleic Acids Res* 2004;33:D29–33. <https://doi.org/10.1093/nar/gki098>.
- [3] Cunningham F, Allen JE, Allen J, Alvarez-Jarreta J, Amode MR, Armean IM, et al. Ensembl 2022. *Nucleic Acids Res* 2022;50:D988–95. <https://doi.org/10.1093/nar/gkab1049>.
- [4] Papanicolaou A, Heckel DG. The GMOD Drupal Bioinformatic Server Framework. *Bioinformatics* 2010;26:3119–24. <https://doi.org/10.1093/bioinformatics/btq599>.
- [5] in Brest (<https://genomes-catalog.ifremer.fr/>), in Roscoff (<https://phaeoexplorer.sb-roscoff.fr/> and <https://rhodoexplorer.sb-roscoff.fr/>) and Rennes (<https://bipaa.genouest.org/> and <https://bbip.genouest.org/>).
- [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. *Sci Data* 2016;3:160018. <https://doi.org/10.1038/sdata.2016.18>.
- [7] <https://gitlab.com/beaur1s/beauris>.
- [8] <https://docs.gitlab.com/ee/ci/>.
- [9] Cormier, A., Durand, P., Noel, C. ORSON: a nextflow workflow for transcriptome and proteome annotation 2021. <https://doi.org/10.48546/WORKFLOWHUB.WORKFLOW.136.1>
- [10] <https://docs.gitlab.com/runner/>.
- [11] Buels R, Yao E, Diesh CM, Hayes RD, Munoz-Torres M, Helt G, et al. JBrowse: a dynamic web platform for genome visualization and analysis. *Genome Biol* 2016;17:66. <https://doi.org/10.1186/s13059-016-0924-1>.
- [12] Lee E, Helt GA, Reese JT, Munoz-Torres MC, Childers CP, Buels RM, et al. Web Apollo: a web-based genomic annotation editing platform. *Genome Biol* 2013;14:R93. <https://doi.org/10.1186/gb-2013-14-8-r93>.
- [13] Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K, et al. BLAST+: architecture and applications. *BMC Bioinformatics* 2009;10:421. <https://doi.org/10.1186/1471-2105-10-421>.
- [14] <https://github.com/abrethead/docker-sf-blast>.
- [15] Holmer R, van Velzen R, Geurts R, Bisseling T, de Ridder D, Smit S. GeneNoteBook, a collaborative notebook for comparative genomics. *Bioinformatics* 2019;35:4779–81. <https://doi.org/10.1093/bioinformatics/btz491>.

FlashTalk 1

Compositional biases promoting self-assembly establish a link between the genome- and the cell-spatial self-organization

Audrey Lapendry¹, Nicolas Fontrodona¹, Audrey Gibert¹ and Didier Auboef¹

¹ Laboratoire de Biologie et Modelisation de la Cellule, Ecole Normale Supérieure de Lyon, CNRS, UMR 5239, Inserm, U1293, Université Claude Bernard Lyon 1, 46 allée d'Italie, F-69364 Lyon, France

Corresponding Author: didier.auboef@inserm.fr

In recent years, significant progress has been made in understanding the role of the genome spatial organization in gene expression regulation [1]. The genome spatial organization can be defined by the fact that genes are not randomly distributed in the nucleus, but are instead organized within more or less dynamical spatial communities or clusters.

To better understand the biological causes and consequences of the genome spatial organization, we first collected different kind of datasets that allow to infer the spatial localization of human genes in the nucleus. For example, we used several human Hi-C datasets that allow to define groups of genes that are in spatial proximity, datasets that allow to define the nuclear radial position of human genes [2], and datasets that allow to define the localization of genes in regards to nuclear sub-compartments such as lamina, speckles and nucleolus [3]. Using these datasets and analyzing gene nucleotide composition, we show that genes that are in close spatial proximity to each other share the same nucleotide composition bias. The composition biases – and thus the physicochemical properties shared between genomic regions – could explain at least in part how they self-assemble within spatial communities. We also show that co-localized genes sharing the same compositional biases have a higher probability of being co-regulated by the same set of transcription factors. We then show that co-localized genes that have the same compositional biases produce RNAs that share the same biases and that are co-regulated by the same set of RNA-binding proteins, such as splicing factors. Then, we show that RNAs produced from the same gene spatial clusters and sharing the same compositional biases produce proteins that themselves share the same amino acid compositional biases. Consequently, proteins whose genes co-localize have the same physicochemical properties. Since the cellular localization of proteins depends on their physicochemical properties, we next show that proteins whose genes co-localize in the nucleus have a higher probability of being part of the same cellular sub-compartments.

Through the analysis of the compositional biases of nucleic acids and proteins – as a proxy of their physicochemical properties – our work uncovers a link between the spatial organization of genes in the nucleus and the spatial organization of their products (i.e., proteins) in the cell. In addition, our work supports a model according to which gene- function and -regulation are two sides of the same coin. Indeed, genes are co-regulated according to their compositional biases and because they share nucleotide composition biases, co-localized and co-regulated genes produce proteins that share the same amino acid composition biases and that have therefore similar biological functions. In a symmetrical manner, proteins that have similar biological functions share the same amino acid composition biases and therefore come from co-localized genes sharing the same nucleotide composition biases and co-regulated by the same factors.

References

- [1] Bouwman, Britta A. M., Nicola Crosetto, et Magda Bienko. The Era of 3D and Spatial Genomics. *Trends in Genetics*: TIG 38 (10): 1062-75, 2022.
- [2] Girelli, Gabriele, Joaquin Custodio, Tomasz Kallas, Federico Agostini, Erik Wernersson, Bastiaan Spanjaard, Ana Mota, et al. GPSeq Reveals the Radial Organization of Chromatin in the Cell Nucleus. *Nature Biotechnology* 38 (10): 1184-93, 2020.
- [3] Wang, Yuchuan, Yang Zhang, Ruochi Zhang, Tom van Schaik, Liguang Zhang, Takayo Sasaki, Daniel Peric-Hupkes, et al. SPIN Reveals Genome-Wide Landscape of Nuclear Compartmentalization. *Genome Biology* 22 (1): 36, 2021.

DLscaff : Deep Learning and Hi-C data for chimeric contigs detection.

Alexis MERGEZ¹, Raphaël MOURAD¹ and Matthias ZYTNIKI¹

¹ MIAT, Toulouse INRAE, 31326 CEDEX, Castanet-Tolosan, France

Corresponding Author: alexis.mergez@inrae.fr

De novo whole genome assembly is a difficult task that aims to reconstruct the genome of an organism with relatively small fragments called reads and no reference as backbones [1]. To achieve this, several algorithms exist but they rely on alignment or overlap between reads or contigs in order to make linkage [1]. This method is error prone due to a number of variables such as sequencing depth, sequencing error or repeated sequences [1]. Errors tend to propagate to following steps, decreasing the overall initial quality of an assembly [2].

Today's projects like the Darwin Tree of Life (DToL) project use a combination of protocol and sequencing techniques to create initial assemblies of high quality but not error-free [3]. Each assembly needs to be curated by hand [2, 3] which is a long task. Hence, the resources and time associated with these projects are great [2, 3].

Invented around 2010, the Hi-C protocol allows us to capture the spatial conformation of the genome [4, 5]. It relies on linking close sequences together, sequencing them, then counting the links between different parts of the genome [6]. One key observation is that sequences that are spatially close are also close within the genome [4]. This makes this protocol very useful for *de novo* assemblies as it helps to build the scaffolds based on linkage correlation between contigs.

One of the typical errors occurring in contigs assembly results in a chimeric contig. It can be defined as a contig that is composed of two unrelated sequences, from different chromosomes for example [7]. They can be joined because of a shared repeated sequence that mislead the assembler [7]. Several tools exist to detect and correct those errors. YaHS seems to be the best current solution and is being used in the DToL project [8]. It works by searching for a position where the number of linkages is lower than a certain threshold. The assumption is that a low count position is representative of a misjoin between two unrelated sequences [8]. This method saw a lot of different implementations with their set of parameters [8, 9].

Our proposed method relies on applying a deep learning computer vision model on the Hi-C contact matrix directly. The principle is to remove the reliance on a threshold or other metrics that can require dataset specific adjustments and rather search for patterns on Hi-C contact matrix. Indeed, chimeric contigs can be detected on Hi-C contact map because of the pattern they induce [8]. Preliminary results seem to indicate that our method outperforms current state of the art methods using a self-supervised model [10]. A lot of work is still required to search for the best model as well as find the best hyper-parameters.

References

- [1] Liao and al. "Current Challenges and Solutions of de Novo Assembly." *Quantitative Biology* 7, no. 2 (2019): 90–109.
- [2] Howe and al. "Significantly Improving the Quality of Genome Assemblies through Curation." *GigaScience* 10, no. 1 (2021): g1aa153.
- [3] The Darwin Tree of Life Project Consortium. "Sequence Locally, Think Globally: The Darwin Tree of Life Project." *Proceedings of the National Academy of Sciences* 119, no. 4 (2022): e2115642118.
- [4] Lieberman-Aiden and al. "Comprehensive Mapping of Long-Range Interactions Reveals Folding Principles of the Human Genome." *Science* 326, no. 5950 (2009): 289–93.
- [5] Belton and al. "Hi-C: A Comprehensive Technique to Capture the Conformation of Genomes." *Methods (San Diego, Calif.)* 58, no. 3 (2012).
- [6] Pal and al. "Hi-C Analysis: From Data Generation to Integration." *Biophysical Reviews* 11, no. 1 (2018): 67–78.
- [7] Pan and al. "Accurate Detection of Chimeric Contigs via Bionano Optical Maps." *Bioinformatics* 35, no. 10 (2019): 1760–62.
- [8] Zhou and al. "YaHS: Yet Another Hi-C Scaffolding Tool." *Bioinformatics* 39, no. 1 (2023): btac808.
- [9] Wang and al. "EndHiC: Assemble Large Contigs into Chromosome-Level Scaffolds Using the Hi-C Links from Contig Ends." *BMC Bioinformatics* 23, no. 1 (2022): 1–19.
- [10] Dwivedi and al. "With a Little Help from My Friends: Nearest-Neighbor Contrastive Learning of Visual Representations." *arXiv*, October 7, 2021.

Exploring the Potential of a Biomimetic Fibronectin Motif to Interact with Type I Collagen for Tissue Regeneration: In Silico and Experimental Analyses

Jad Eid*¹

¹EBInnov® – Ecole de Biologie Industrielle – France

Résumé

Collagen and Fibronectin (FN) are two vital components of the extracellular matrix that play a crucial role in regulating cellular behavior and matrix remodeling through their interaction (1). The fibronectin collagen binding domain (F-CBD) has garnered interest in tissue engineering and regenerative medicine applications due to these functions (2,3). In pursuit of developing an innovative bio-adhesive motif for these applications, we have designed a new chimeric protein containing FN-collagen and cell binding domains (4). The protein comprises the F-CBD and cellular binding domains (PHSRN and RGD amino acids) of the 9th and 10th fibronectin type 3 (FNIII9 and FNIII10) modules, which we have cloned and produced at the EBI laboratory. However, questions remain about the affinity interaction of the collagen-chimera complex and the possible conformations of the collagen protein within the chimera.

Therefore, our project aims to bioproduce the chimeric protein and perform collagen binding tests to study the chimera-collagen interaction. Additionally, we utilized several bioinformatics tools, including AlphaFold and I-TASSER, to predict the three-dimensional structure of the chimera and explore the homologues using BLAST. We also used AutoDock Vina to study the docking of collagen in the chimera. These tools allowed us to investigate the possible conformations of the collagen protein in the chimera and to assess the affinity interaction of the collagen-chimera complex. We have successfully produced the chimeric protein, and our experimental results are consistent with our bioinformatics predictions, which illustrate the high affinity between collagen and the chimera. These findings present many opportunities for the application of this chimera in tissue engineering and regenerative medicine.

(1) Patten J, Wang K. Fibronectin in development and wound healing. *Adv Drug Deliv Rev.* 2021 Mar;170:353-368. doi: 10.1016/j.addr.2020.09.005. Epub 2020 Sep 19. PMID: 32961203.

(2) Parisi L, Toffoli A, Ghezzi B, Mozzoni B, Lumetti S, Macaluso GM. A glance on the role of fibronectin in controlling cell response at biomaterial interface. *Jpn Dent Sci Rev.* 2020 Dec;56(1):50-55. doi: 10.1016/j.jdsr.2019.11.002. Epub 2019 Dec 18. PMID: 31890058; PMCID: PMC6928270.

(3) Pankov R, Yamada KM. Fibronectin at a glance. *J Cell Sci.* 2002 Oct 15;115(Pt 20):3861-3. doi: 10.1242/jcs.00059. PMID: 12244123.

(4) Ben Abla A, Boeuf G, Elmarjou A, Dridi C, Poirier F, Changotade S, Lutowski D, Elm'selmi A. Engineering of Bio-Adhesive Ligand Containing Recombinant RGD and PHSRN Fibronectin Cell-Binding Domains in Fusion with a Colored Multi Affinity Tag: Simple Approach for Fragment Study from Expression to Adsorption. *Int J Mol Sci.* 2021 Jul 8;22(14):7362. doi: 10.3390/ijms22147362. PMID: 34298982; PMCID: PMC8303147.

*Intervenant

Mots-Clés: Collagen, Fibronectin, Extracellular matrix, Biomimetic scaffold, Chimeric protein, Genetic engineering, Homology modeling, 3D structure prediction, Docking, AutoDock Vina, conformations, affinity, Bioinformatics, Tissue engineering, Regenerative medicine

Genes encoding teleost orthologs of human signal transduction proteins remain in duplicate or in triplicate more frequently than the whole genome

Floriane Picolo^{*1}, Benoît Piégu¹, and Philippe Monget^{†1}

¹Physiologie de la reproduction et des comportements [Nouzilly] – Institut Français du Cheval et de l'Equitation [Saumur], Université de Tours, Centre National de la Recherche Scientifique, Institut National de Recherche pour l'Agriculture, l'Alimentation et l'Environnement – France

Résumé

Introduction

Cell signaling involves many proteins, many of which belong to families of related proteins, and these proteins together display a huge number of interactions. One of the events that led to the creation of new genes was whole genome duplication (WGD), which made some major innovations possible. In addition to the two WGDs that happened in vertebrate genomes, teleost genomes underwent a third WGD after separation from the lineage leading to holostei. Moreover, a fourth WGD also occurred independently in salmonids (100MYA) and carps (10MYA) (1,2). One of the major assumptions of the preservation of duplicated genes is a dose effect (3). Indeed, this postulate is based on the idea that the number of copies of a gene is influenced by the dosage of the products with which it interacts (4).

Materials and Methods

In the present work, we have studied in 63 teleost species whether the orthologs of human genes involved in each of the 47 signaling pathways (HGSP) remain more frequently in duplicates, triplicates or returned more frequently as singleton than the whole genome.

Results

Our results show these genes remain more frequently in duplicate and in triplicate in teleost of the 3WGD and 4WGD group, respectively. We obtained similar results for teleost orthologs of genes encoding ligand/membrane receptor pairs (5). Moreover, by examining pairs of interacting genes products in terms of conserved copy numbers, we show that a majority of 1:1 and 1:2 proportion/stoichiometry, and of 2:2 and 2:4 proportion was observed in the 3WGD (between 54% and 60%) and 4WGD (30%) group, respectively. In both groups, the 0:n proportion was observed at a mean of approximately 10%, some pseudogenes being found in the concerned genomes. Finally, the proportions were very different between the studied pathways. The n:n (i.e. same) proportion concerns from 20% to 65% of the interactions, depending on the pathways, the n:m (i.e. different) proportion from 34% to 70%.

*Intervenant

†Auteur correspondant: philippe.monget@inrae.fr

Among the n:n proportion, the 1:1 ratios is the more represented (25.8%) and among the n:m ratios, the 1:2 is the more represented (25.0%). An absence of gene loss was observed for the JAK-STAT, FoxO and Glucagon pathways. Overall, these results show that the teleost genes orthologs of HGSP remain more in duplicate (3WGD) and in triplicate (4WGD) than the whole genome, some genes being lost, and the proportions being not always maintained.

References

I. Braasch et J. H. Postlethwait, " Polyploidy in Fish and the Teleost Genome Duplication ", in *Polyploidy and Genome Evolution*, P. S. Soltis et D. E. Soltis, Éd., Berlin, Heidelberg: Springer, 2012, p. 341-383. doi: 10.1007/978-3-642-31442-1_17. S. Lien *et al.*, " The Atlantic salmon genome provides insights into rediploidization ", *Nature*, vol. 533, no 7602, p. 200-205, mai 2016, doi: 10.1038/nature17164.

- B. Papp, C. Pál, et L. D. Hurst, " Dosage sensitivity and the evolution of gene families in yeast ", *Nature*, vol. 424, no 6945, Art. no 6945, juill. 2003, doi: 10.1038/nature01771.

- M. Freeling et B. C. Thomas, " Gene-balanced duplications, like tetraploidy, provide predictable drive to increase morphological complexity ", *Genome Res.*, vol. 16, no 7, p. 805-814, juill. 2006, doi: 10.1101/gr.3681406. A. Grandchamp, B. Piégu, et P. Monget, " Genes encoding teleost fish ligands and associated receptors remained in duplicate more frequently than the rest of the genome ", *Genome Biol. Evol.*, avr. 2019, doi: 10.1093/gbe/evz078.

Mots-Clés: WGD, duplication, genes, teleosts, signaling pathway

Handling confounding factors in analyzing the transcriptomic data from Chernobyl tree frogs

Elen GOUJON^{1,3}, Olivier ARMANT², Jean-Marc BONZOM², Arthur TENENHAUS³ and Imène GARALI¹

¹ Institut de radioprotection et de sûreté nucléaire, PSE-SANTE/SESANE/LRTOX, BP 17, 92260, Fontenay-aux-Roses, France

² Institut de radioprotection et de sûreté nucléaire, PSE-ENV/SRTE/LECO, BP 3, 13115, Saint-Paul-lez-Durance, France

³ Université Paris-Saclay, CNRS, CentraleSupélec, Laboratoire des signaux et systèmes (L2S), 3 rue Joliot-Curie, 91190, Gif-sur-Yvette, France

Corresponding Author: imene.garalizineddine@irsn.fr

Abstract

More than 30 years after the nuclear power plant accident in Chernobyl, wildlife in the Chernobyl Exclusion Zone (CEZ) is still chronically exposed to low doses of ionizing radiation. In 2018, populations of the Eastern tree frog *Hyla orientalis* were sampled inside the CEZ across a gradient of radiocontamination and in nearby control sites [1]. Our research project aims at developing methods for multi-omics data integration to advance the understanding of the consequences of chronic exposure to low-dose radiation. The first axis of this study focuses on the exploration of transcriptomic data in relation to dosimetry. The analysis of RNA-seq data could reveal radiation-specific molecular signatures; however, this can only be accomplished if confounding factors are properly accounted for. Here, we compare batch effect removal strategies using the transcriptomic data obtained from 87 tree frogs. We highlight the strengths and weaknesses of the following methods: the ComBat-seq method dedicated for RNA-seq count data [2], linear regression on the batch with residuals extraction, and an integrated approach with mixOmics MINT [3]. The geographical site of origin of the frogs is a confounding factor for the analysis, and its complete confoundedness with the dose makes its handling challenging. Indeed, the different methods we used were brought to their limits: correcting or accounting for the batch effect cuts out the biological information related to the dose. The site factor encompasses hidden information, including the local environment and available food sources, the day of capture, and the genetic diversity between the populations of tree frogs. In a second approach, we performed hierarchical clustering on the genetic distance matrix to summarize the genetic diversity into a batch factor. Using the methods previously mentioned, we were able to integrate genetic diversity into the analysis and bring to light the effect of the ionizing radiation dose in the transcriptomic data.

References

1. Clément Car *et al.* Unusual evolution of tree frog populations in the Chernobyl exclusion zone. *Evolutionary Applications*, (15):203-219, 2022.
2. Yuqing Zhang *et al.* ComBat-seq: batch effect adjustment for RNA-seq count data. *NAR Genomics and Bioinformatics*, (3):lqaa078, 2020.
3. Florian Rohart *et al.* MINT: a multivariate integrative method to identify reproducible molecular signatures across independent experiments and platforms. *BMC Bioinformatics*, (18):128, 2017.

THEMA: Identification of molecular mechanisms by which Tumor Heterogeneity influences disease outcome: high-dimension Mediation Analysis to link causes and consequences

F. PITTION, M. ESTAVOYER, B. JUMENTIER, O. FRANÇOIS and M. RICHARD
MAGe team, TIMC, CNRS UMR 5525, Grenoble, France

Corresponding author: florence.pittion@univ-grenoble-alpes.fr

Poster Abstract

As tumors are dynamic heterogeneous ecosystems, it is essential to study the underlying heterogeneity for a better understanding of cancer biology. Currently, there is no framework to infer the causal mechanisms underlying the effect of tumor heterogeneity on disease outcome. Heterogeneity and composition of the tumor has a major impact on cancer cells growth, division, resistance and metastasis [1]. We want to test the extent to which the effect of tumor heterogeneity on disease outcome is explained (or not) by the molecular features of the tumor (i.e. gene expression and DNA methylation, an epigenetic mark regulating the gene expression). Our goal is to develop a **new multimodal high-dimension mediation analysis framework** to unravel this causal links.

THEMA project will concentrate on one critical scenario: the development of pancreatic ductal adenocarcinoma (PDAC). Incidence of this cancer increases regularly in Western countries and is expected to become the second leading cause of cancer-related mortality in 2025. Our collaborators [2] recently demonstrated that PDACs are highly heterogeneous cancers, with an abundant differentiated stroma associated with prognostic relevance.

First, we will study the indirect effect of **DNA methylation (DNAm)** in the pathway that links tumor heterogeneity and disease outcome [3]. DNAm is a well-studied epigenetic mechanism that contributes to cell type differentiation through the control of gene expression. We will then extend this model to study the indirect effect of **gene expression** in the same pathway. Third, we will develop a statistical framework to perform **multimodal** high-dimension mediation analysis and question the relationship between the identified mediators (gene expression and DNAm). Finally, we will test how the exposure to an external environmental cue such as **therapeutic treatment** affect the identified indirect effects.

We expected that THEMA will: i) identify for the first time molecular mediators of tumor heterogeneity, both at the DNA methylation and gene expression level; ii) improves understanding of carcinogenesis; iii) offers great perspectives in the development of new biomarkers and personalized therapeutic treatments, with combined mediation analysis of external environmental cues.

In this poster I will present context, goals and methodology of my doctoral project. I will also present mediation analysis simple approach results and conditional simulation first results (based on the existing methylation matrix).

References

- [1] Andriy Marusyk, Michalina Janiszewska, and Kornelia Polyak. Intratumor heterogeneity: the Rosetta stone of therapy resistance. *Cancer cell*, 37(4):471–484, April 2020.
- [2] Francesco Puleo, Rémy Nicolle, Yuna Blum, et al. Stratification of Pancreatic Ductal Adenocarcinomas Based on Tumor and Microenvironment Features. *Gastroenterology*, 155(6):1999–2013.e3, December 2018.
- [3] Peter A. Jones. Functions of DNA methylation: islands, start sites, gene bodies and beyond. *Nature Reviews Genetics*, 13(7):484–492, Jul 2012.

Recent introduction of *Angiostrongylus cantonensis* and its intermediate host *Achatina fulica* in Guadeloupe evidenced by phylogenetic analyses

Gélixa Gamiette*¹

¹Institut Pasteur de la Guadeloupe – Guadeloupe

Résumé

Angiostrongylus cantonensis (rat lungworm) is the main pathogen responsible for eosinophilic meningitis in humans. Its intermediate host *Achatina fulica* invaded many countries around the world before appearing in the West Indies in the late 1980s. In our departments, the first cases of human angiostrongylosis were reported in Martinique, Guadeloupe and French Guiana in 2002, 2013 and 2017 respectively. In order to have a better knowledge of angiostrongylosis in Guadeloupe, particularly its geographical origin and mode of introduction, this study proposes a molecular characterization of adult worms of *A. cantonensis* and its intermediate host *A. fulica*.

Mots-Clés: *Angiostrongylus cantonensis*, *Achatina fulica*, Guadeloupe, phylogeny, cytochrome C, cytochrome B, rRNA 16s

*Intervenant

End of the beginning: telomeres are only at one side of the chromosomes in the nematode *Meloidogyne incognita*.

Ana-Paula ZOTTA MOTA¹, Georgios D KOUTSOVOULOS¹, Laetitia PERFUS-BARBECH¹, Evelin DESPOT-SLADE², Karine LABADIE³, Jean-Marc AURY⁴, Karine ROBBE-SERMESANT¹, Marc BAILLY-BECHET¹, Caroline BELSER⁴, Arthur PÉRÉ¹, Corinne RANCUREL¹, Djampa K KOZLOWSKI^{1,5}, Rahim HASSANALY-GOULAMHOUSSEN¹, Martine DA ROCHA¹, Benjamin NOËL⁴, Nevenka MEŠTROVIĆ², Patrick WINCKER⁴ & Etienne GJ DANCHIN^{1*}

¹ Institut Sophia Agrobiotech, INRAE, Université Côte d'Azur, CNRS, 400 routes des Chappes, 06903 Sophia-Antipolis, France

² Division of Molecular Biology, Ruđer Bošković Institute, Bijenička cesta 54, 10000, Zagreb, Croatia

³ Genoscope, Institut François Jacob, CEA, CNRS, Univ Evry, Université Paris-Saclay, 91057 Evry, France

⁴ Génomique Métabolique, Genoscope, Institut François Jacob, CEA, CNRS, Univ Evry, Université Paris-Saclay, 91057 Evry, France

⁵ Université Côte d'Azur, Center of Modeling, Simulation, and Interactions, 28 avenue Valrose, 06000 Nice, France

Corresponding Author: karine.robbe-sermesant@inrae.fr

Assembling the genomes of polyploid hybrid species is a challenge, due to variable degrees of divergence between the genome copies. Unlike in polyploid hybrids, classical sexually-reproducing species have diploid genomes with two highly similar chromosome sets inherited from their parents. However, for the root-knot nematode *Meloidogyne incognita*, it is hypothesized that three chromosome sets with various degrees of sequence divergence compose its triploid (A'A"B) genome.

We have generated genomic data using long-reads from ONT technology for *M. incognita*. Using NECAT we assembled the genome in 291 contigs which we further polished with both ONT long reads and Illumina short reads. This allowed resolving and assembling highly repetitive regions of the *M. incognita* genome for the first time. Thanks to this new assembly, we could investigate the telomere sequences of *M. incognita* [1]. Against all odds, we could find neither the telomerase genes nor the canonical telomeric repeat sequence of the model nematode *C. elegans*. However, at the end of some *M. incognita* contigs, we found an enrichment of a composite G-rich repeat. We compared this new repetitive sequence against all nematode species with a sequenced genome, and apart from partial matches in closely related *Meloidogyne* species, the repeat was absent from the rest of the nematodes. Using fluorescent in-situ hybridization we confirmed the repeat had a telomeric localization but only at one extremity of the chromosomes.

The discovery of a new kind of telomeric repeat in these species highlights the evolutionary diversity of chromosome protection systems despite their central roles and opens new perspectives towards the development of more specific control methods against these pests.

References

1. Ana-Paula Zotta Mota. Unzipped assemblies of polyploid root-knot nematode genomes reveal new kinds of unilateral composite telomeric repeats. bioRxiv, 2023

CellFromSpace: A versatile tool for spatial transcriptomic data analysis through reference-free deconvolution and guided cell type/activity annotation

Corentin THUILLIEZ¹, Maria Eugenia MARQUES DA COSTA¹, Pierre KHNEISSER², Jean-Yves SCOAZEC²,
Nathalie GASPARD^{1,3}, Gael MOQUIN-BEAUDRY¹ and Antonin MARCHAIS^{1,3}

¹ INSERM U1015, Gustave Roussy Cancer Campus, Université Paris-Saclay, 114 Rue Edouard Vaillant, 94805, Villejuif, France

² Department of Pediatric and Adolescent Oncology, Gustave Roussy Cancer Campus, 94805, 114 Rue Edouard Vaillant, 94805, Villejuif, France

³ Department of Medical Biology and Pathology, Gustave Roussy Cancer Campus, Université Paris-Saclay, 94805, 114 Rue Edouard Vaillant, 94805, Villejuif, France

Corresponding Author: corentin.thuilliez@gustaveroussy.fr

Spatial transcriptomics approaches have recently emerged as some of the most promising technologies to analyze spatial distribution of cell types, and gene expression within tissues. These technologies are mainly divided in two categories: transcript level panel-based high throughput in situ hybridization/sequencing approaches (Vizgen Merscope, Nanostring CosMX [1], 10X Xenium [2]) and spatially barcoded next generation sequencing-based approaches (10X Visium, Slide-seqV2, Stereo-seq) [3]. The latter enable the detection of thousands of transcripts on tissue sections at subcellular to quasicellular resolution. However, this results in highly dimensional, sparse, spatially distributed data which can be highly demanding on computational resources. 10X's Visium, currently the most widespread technology by publication metrics [3], outputs manageable data size at the cost of lower resolution, compared to other techniques, with 55µm diameter spots typically encompassing 1-20 cells. Therefore, a deconvolution step is often required to gain insight into the mixture of cells populating each spot: these deconvolution methods are often scRNA-seq reference based, with known drawbacks such as the loss of information and the necessity of a high-quality reference datasets.

Here, we propose a new method named CellFromSpace (CFS), based on independent component analysis, enabling an unsupervised deconvolution of Visium data. We developed a package associated with a shiny tool that accelerates the annotation of deconvolved signal by biologists and/or clinicians. Thanks to the visual interface of the shiny tool, data representations and annotation can easily be achieved using the generated heatmaps, pathway analyses, spatial plotting and scatter pie charts. Spots can then be clustered and projected on UMAP; marker genes of clusters can be determined as well as ligand/receptors and cell type interactions. Results can easily be subsetted and output from the tool for further analysis by bioinformaticians. To evaluate the efficiency of CFS, we analyzed the visium dataset available from 10x Genomics Spatial Gene Expression dataset. Visium fresh frozen and FFPE samples of adult mouse brain and human tumors were analyzed.

CFS was able to infer the cell composition of the mouse brain with high fidelity of the different layers of the brain but also detected signatures of diffuse cells such as astrocytes and microglial cells. Furthermore, CFS identified spot composition in cell types and activities within heterogeneous cancer tissues matching the pathologist annotations. Our method allows for the identification of different phenotypes within a tumor and identify peripheric and infiltrating stromal signatures such as immune infiltrates. CFS also enable to subset particular cell population to analyze them separately only using the corresponding independent components and spots.

In conclusion, CFS is a tool that offers the possibility to thoroughly analyze and easily interpret results from NGS-based spatial transcriptomics analysis in absence of single cell dataset.

References

1. Shanshan He, and al. High-Plex Multiomic Analysis in FFPE Tissue at Single-Cellular and Subcellular Resolution by Spatial Molecular Imaging. *bioRxiv* 2021.11.03.467020
2. Amanda Janesick, and al., High resolution mapping of the breast cancer tumor microenvironment using integrated single cell, spatial and in situ analysis of FFPE tissue, *bioRxiv* 2022.10.06.510405
3. Moses, L., Pachter, L. Museum of spatial transcriptomics. *Nat Methods* 19, 534–546 (2022)

Méthodes spatiales pour l'analyse de données génétiques

M. GUIVARCH¹, G. LE FOLGOC¹, G. MARENNE¹, T. LUDWIG^{1,2}, I. ALVES³, C. DINA³, R. REDON³, J. M. SEBAOUN⁴, L. GRESSIN⁴, J. C. BEAUDOIN⁴, V. MOREL⁴, B. FIN⁵, C. BESSE⁵, R. OLASO⁵, D. BACQ⁵, V. MEYER⁵, F. SANDRON⁵, A. FERRANE⁶, A. BOLAND⁵, H. BLANCHÉ⁴, M. ZINS⁷, J. F. DELEUZE^{4,5}, E. GÉNIN^{1,2}, A. F. HERZIG¹, A. SAINT PIERRE¹

1 Univ. Brest, Inserm, EFS, UMR 1078, GGB, IBSAM, F-29200, Brest, France

2 CHU Brest, F-29200 Brest, France

3 Nantes Université, CHU Nantes, CNRS, INSERM, l'institut du thorax, F-44000 Nantes, France

4 Fondation Jean Dausset, CEPH, Paris, France

5 Université Paris-Saclay, CEA, Centre National de Recherche en Génomique Humaine (CNRGH), F-91057, Evry, France

6 Institut de Santé Publique, Pôle de recherche clinique, INSERM, F-75013 Paris, France

7 Université de Paris Cité, Université Paris-Saclay, UVSQ, Inserm, Cohortes Epidémiologiques en Population, UMS 11, F-94807, Villejuif, France.

Corresponding Author: mael.guivarch@univ-brest.fr

Le projet POPGEN couvre l'ensemble du territoire métropolitain et regroupe les informations génétiques et géographiques précises de plus de 9772 volontaires issus de la cohorte Constances. Les individus ont été sélectionnés sur la proximité entre les lieux de naissance de leurs grands-parents. Il s'agit donc du plus grand projet en France qui fasse le lien entre géographie et génétique.

Actuellement, la plupart des méthodes permettant de détecter les structures de population n'intègrent pas directement la géographie [1]. Lorsque l'information est disponible, elle est généralement considérée a posteriori pour interpréter les patterns de diversité génétique observés.

On s'appuie ici sur les données du projet POPGEN pour présenter des résultats de clustering génétique qui dévoilent une stratification géographique fine de la population [2,3]. Les résultats de partitionnement font notamment ressurgir clairement des barrières, topographiques ou culturelles, au niveau du Pays basque, du Nord, de la Bretagne et des Vosges.

A partir de ces constatations, on propose de développer des méthodes basées sur l'utilisation des informations apportées par les données spatiales pour analyser plus finement des schémas de stratification génétique et pour mieux comprendre les différences phénotypiques observées dans les différentes régions du territoire métropolitain.

On comparera en particulier deux types d'approches spatiales : d'une part, les méthodes géostatistiques [4,1], telles que le krigeage, qui intègrent des informations géographiques dans l'analyse, et d'autre part les méthodes non-géostatistiques, plus classiquement utilisées en génétique des populations et qui prennent uniquement en compte les distances génétiques entre individus. On montrera notamment que ces deux types d'approches sont complémentaires.

Acknowledgements

The POPGEN project (French Ministry of Research - PFMG2025)

References

- 1- DARLU, Pierre. Les représentations géographiques de la diversité biologique dans l'espèce humaine. *L'Espace géographique*, 1997, p. 341-353
- 2- SAINT PIERRE, Aude, GIEMZA, Joanna, ALVES, Isabel, et al. The genetic history of France. *European Journal of Human Genetics*, 2020, vol. 28, no 7, p. 988. SAINT PIERRE, Aude, GIEMZA, Joanna, ALVES, Isabel, et al. Correction: The genetic history of France. *European Journal of Human Genetics*, 2020, vol. 28, no 7, p. 988.
- 3- HERZIG, Anthony F., VELO-SUÁREZ, Lourdes, FREX CONSORTIUM, et al. Can imputation in a European country be improved by local reference panels? The example of France. *bioRxiv*, 2022, p. 2022.02.17.480829.
- 4-MATHERON, Georges. Principles of geostatistics. *Economic geology*, 1963, vol. 58, no 8, p. 1246-1266.

FlashTalk 2

GRUPS-rs, a high-performance ancient DNA genetic relatedness estimation software relying on pedigree simulations.

Maël Lefeuvre¹, Michaël Martin², Flora Jay³, Marie-Claude Marsolier^{1,4†} and Céline Bon^{1†}

1. UMR 7206 - Éco-Anthropologie (EA), Muséum National d'Histoire Naturelle, CNRS, Université Paris Cité, 75016-Paris, France.
2. Department of Natural History, NTNU University Museum, Norwegian University of Science and Technology (NTNU), Trondheim, Norway.
3. Laboratoire Interdisciplinaire des Sciences du Numérique, CNRS, INRIA, Université Paris-Saclay, 91400-Orsay, France
4. CEA/DRF, I2BC/UMR 9198, SBIGeM, Gif-sur-Yvette, France.

Corresponding Author: mael.lefeuvre@mnhn.fr

The study of fine-grain genetic kinship ties (parents, siblings, cousins, etc.) from ancient remains is now gaining significant interest and prominence within the field of paleogenetics, as a means of deciphering the social organization of past societies^[1,2]. However, despite sustained research efforts, kinship analyses are in practice often quite difficult to apply within paleogenetic studies^[1,3], and may carry a high degree of uncertainty in the results they provide, especially when working with lowly covered, highly degraded samples – as is typically the case in the context of ancient DNA studies – or studying poorly characterized populations. To overcome these challenges, most of the methods dedicated to kinship estimation between ancient DNA samples either refrain from inferring kinship past the second-degree of relatedness^[4,5] (e.g.: half-siblings), and/or rely on the use of a *cohort* of individuals to obtain a satisfactory statistical significance^[4,6,7]. Thus, the current state of the art remains intrinsically limited when attempting to estimate kinship on a *small* number of individuals, or when trying to detect more distant relationships (e.g.: cousins).

Here, we present an update and complete reimplementations of "GRUPS" (*Get Relatedness Using Pedigree Simulations*): an ancient DNA kinship estimation software based on the methods originally developed in (Martin D. et al – 2017)^[8]. GRUPS both computes an estimate of relatedness from randomly sampled pseudo-haploidized variant calls, and leverages high-definition pedigree simulations to bypass the use of a cohort of individuals, making this method especially suitable when attempting to perform kinship analysis on a single pair of low-coverage individuals. We highlight that GRUPS can provide with a sufficient statistical significance to estimate genetic relatedness past the second degree, while taking into account contamination and sequencing error estimates when simulating individuals. Finally, our updated method, "GRUPS-rs" offers an estimated 2200-fold speed-up in runtime performance compared to its predecessor – allowing the joint estimation of kinship between dozens of individuals in a matter of minutes – and is now bundled with a user-friendly Shiny interface, in which users can interactively visualize their results.

References

- [1] Vai S, Amorim CEG, Lari M, Caramelli D. Kinship Determination in Archeological Contexts Through DNA Analysis. *Front Ecol Evol.* 2020;8. doi:10.3389/fevo.2020.00083
- [2] Fowler C, Olalde I, Cummings V, et al. A high-resolution picture of kinship practices in an Early Neolithic tomb. *Nature.* 2021;601(7894):584-587. doi:10.1038/s41586-021-04241-4
- [3] Marsh WA, Brace S, Barnes I. Inferring biological kinship in ancient datasets: comparing the response of ancient DNA-specific software packages to low coverage data. *BMC Genomics.* 2023;24(1). doi:10.1186/s12864-023-09198-4
- [4] Monroy Kuhn JM, Jakobsson M, Günther T. Estimating genetic kin relationships in prehistoric populations. Calafell F, ed. *PLoS ONE.* 2018;13(4):e0195491. doi:10.1371/journal.pone.0195491
- [5] Fernandes DM, Cheronet O, Gelabert P, Pinhasi R. TKGWV2: an ancient DNA relatedness pipeline for ultra-low coverage whole genome shotgun data. *Sci Rep.* 2021;11(1). doi:10.1038/s41598-021-00581-3
- [6] Kennett DJ, Plog S, George RJ, et al. Archaeogenomic evidence reveals prehistoric matrilineal dynasty. *Nat Commun.* 2017;8(1). doi:10.1038/ncomms14115
- [7] Popli D, Peyrégne S, Peter BM. KIN: a method to infer relatedness from low-coverage ancient DNA. *Genome Biol.* 2023;24(1). doi:10.1186/s13059-023-02847-7
- [8] Martin, MD, Jay, F, Castellano, S, Slatkin, M. (2017) Determination of genetic relatedness from low-coverage human genome sequences using pedigree simulations. *Mol Ecol.* 2017; 26: 4145– 4157. <https://doi.org/10.1111/mec.14188>

L'impact des outils d'assemblage sur le typage des pathogènes bactériens

Déborah MERDA¹, Maroua SAYEB², Marina CAVAIUOLO³, Claire YVON², Mathilde BONIS³, Virginie CHESNAIS¹

¹ Université Paris Est, ANSES, unité SPAAD, Maisons-Alfort location, F-94701 Maisons-Alfort, France

² Université Paris Est, ANSES, unité SEL, Maisons-Alfort location, F-94701 Maisons-Alfort, France

³ Université Paris Est, ANSES, unité SBCL, Maisons-Alfort location, F-94701 Maisons-Alfort, France

Corresponding Author: deborah.merda@anses.fr

L'émergence de maladies dans les troupeaux ainsi que les toxi-infections alimentaire collectives sont des enjeux majeurs en santé publique mais également d'un point de vue économique. C'est pourquoi des systèmes de surveillance d'agents pathogènes sont mis en place, au niveau national et au niveau européen, c'est le cas par exemple pour *Salmonella* [1] et *Listeria monocytogenes* [2]. La caractérisation de ces agent pathogènes repose sur des approches de typage moléculaire qui sont principalement réalisées à partir de données de séquençage WGS (Whole genome sequencing). Les données obtenues par séquençage doivent être assemblées afin de reconstituer les génomes bactériens permettant leur caractérisation à l'échelle moléculaire. L'assemblage représente donc une étape cruciale pour la qualité des résultats de typage, puisqu'avec des short reads, il est peu probable d'obtenir des génomes bactériens complets, mais plutôt des génomes fractionnés en plusieurs contigs. Dans cette étude, nous avons donc réalisés des séquençage *in-silico* de 24 agents pathogènes d'intérêts afin d'évaluer l'impact de l'assemblage sur les résultats de typage. Pour chaque agent pathogène, la simulation de données de séquençage a été réalisée à partir de 5 génomes ou plus, disponibles sur les banques de données publiques. Les reads ont été simulées selon plusieurs valeurs de profondeur de séquençage (25X, 50X, 100X et 150X) et de qualité (phred score supérieur à 30 et phred score inférieur à 30), avec l'outils art [4]. Pour chaque génome, trois jeux de données ont été simulés de manière indépendante. Ces jeux de données ont ensuite été assemblés par 3 outils spades [4], unicycler [5] et shovill [6] en triplicates afin de tester la reproductibilité de ces méthodes. Le typage des agents pathogènes a enfin été réalisé en utilisant les méthodes de MLST et cgMLST avec l'outil mlst [7] et chewBBACA [8]. Les résultats de MLST ne semblent pas impactés par les différentes méthodes d'assemblage, tandis qu'une variabilité a été observée sur les résultats de cgMLST. Cette variabilité dans les résultats cgMLST semblent principalement liés à la composition intrinsèque des génomes bactériens. Les génomes le long desquels il y a une variation importante du pourcentage en GC présentent des résultats d'assemblage non reproductibles ce qui impacte les résultats de cgMLST. Cependant, l'outil unicycler semble donner des résultats plus répétables pour ces agents pathogènes. En conclusion, il est nécessaire de pouvoir prédire ce niveau de variabilité et d'harmoniser les pipelines d'analyses avant l'investigation d'une épidémie afin de pouvoir déterminer le polymorphisme induit par l'évolution des agents pathogènes et le polymorphisme induit par les algorithmes d'assemblage.

1. Bala Swaminathan, Timothy J Barrett, Patricia Fields, Surveillance for Human *Salmonella* Infections in the United States, *Journal of AOAC INTERNATIONAL*, (89), 553–559, 2006.
2. de Valk, H and Jacquet, C and Goulet, V and Vaillant, V and Perra, A and Simon, F and Desenclos, J C and Martin, P, Surveillance of listeria infections in Europe, *Eurosurveillance*, 10, 572, 2005
3. Bankevich A, Nurk S, Antipov D, Gurevich AA, Dvorkin M, Kulikov AS, Lesin VM, Nikolenko SI, Pham S, Prjibelski AD, Pyshkin AV, Sirotkin AV, Vyahhi N, Tesler G, Alekseyev MA, Pevzner PA. SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. *J Comput Biol*. 2012;1:455-77, 2012

4. Weichun Huang, Leping Li, Jason R Myers, and Gabor T Marth. ART: a next-generation sequencing read simulator, *Bioinformatics* 28, 593-594, 2012
5. Wick RR, Judd LM, Gorrie CL, Holt KE. Unicycler: Resolving bacterial genome assemblies from short and long sequencing reads. *PLoS Comput Biol.* 8;13(6):e1005595, 2017.
6. Seeman, T. GitHub - tseemann/shovill: Assemble bacterial isolate genomes from Illumina paired-end reads, 2020
7. Seeman, T (2012) GitHub - tseemann/mlst: Scan contig files against PubMLST typing schemes
8. Silva M, Machado MP, Silva DN, Rossi M, Moran-Gilad J, Santos S, Ramirez M, Carriço JA. chewBBACA: A complete suite for gene-by-gene schema creation and strain identification. *Microb Genom* 4:000166, 2018.

Modelling of 3D structures of aminoacylases from *Streptomyces ambofaciens* and molecular rules for their specificity and their regioselectivity in lysine N-acylation

Laureline GENNESSEAU¹, Yann GUIAVARC'H¹ and Catherine HUMEAU¹

¹ Reaction and Chemical Engineering Laboratory, 1 Rue Grandville, F-54000, Nancy, France

Corresponding Author: laureline.gennesseaux@univ-lorraine.fr

The N-acylation reaction leads to the addition of an acyl group to the amine function of a molecule which can be an amino acid or a peptide leading to acylated derivatives that have a high application potential because of their technofunctional properties or their biological activities. This reaction can be catalyzed by aminoacylases. Four genes of *Streptomyces ambofaciens*, presenting a very high percentage of homology with the gene sequences of aminoacylases from *Streptomyces mobaraensis* were identified. The construction of deletion mutants of *S. ambofaciens* confirmed the activity of the enzymes encoded by the four genes and revealed the particular interest of two of them, SamAA and SamELA that catalyze the acylation of the α and the ϵ amine function of lysine respectively [1]. A numerical approach was implemented to acquire knowledge on the 3D structure and the catalytic mechanism of SamAA and SamELA which remain unknown until now.

The templates that are defined as proteins having high percentages of identity and similarity with the target sequence and whose structures are known were identified using BLASTp and Swiss Model's "Search for templates" tools. The structures of PDB entries 1Q7L, 5VO3, 4PPZ, 4RUH and 7LGP were identified as SamAA templates with identity percentages between 17 and 33%. Based on these templates, homology models of the 3D structure were built using Modeller from the Discovery Studio suite, Swiss Model and AlphaFold2 and were compared. The best model according to energy calculations and the conservation of key residues was shown to be that obtained by ColabFold. Based on literature survey relating to the templates [2], metal ions were added and structural characteristics were determined: SamAA is a M20 family peptidase and is reasonably supposed to be under a homodimeric form with two zinc atoms per subunit. Residues His88, Asp120, Glu155, Glu182 and His416 are assumed to be responsible for the binding of these ions. Catalytic residues were identified, as Asp90 and Glu154, that would act as a key acid/base residue in the catalytic mechanism.

Using BLASTp and Swiss Model's "Search for templates" tools, the structures of PDB entries 3ICJ, 3IGH, 5YZA, 1RJQ and 6SJ0 were identified as SamELA templates with identity percentages between 12 and 28%. Based on the same strategy as described before, SamELA homology models were built, then, metal ions were added and structural characteristics were determined [3]: SamELA is assumed to be a monomeric protein with two zinc atoms per subunit linked to the residues His71, His73, Asp265, His316, Asp417 and His351. SamELA is not an M20 peptidase like SamAA.

Solvation and energy minimization were applied to the retained models then, first docking simulations were performed aiming to determine the most favourable binding modes with lysine and the key residues and interactions that are responsible for the selectivity of SamAA and SamELA towards the lysine α and ϵ amino group, respectively.

References

1. Lena Dettori, Florent Ferrari, Xavier Framboisier, Cedric Paris, Yann Guiavarc'H, Laurence Hôtel, Arnaud Aymes, Pierre Leblond, Catherine Humeau, Romain Kapel, Isabelle Chevalot, Bertrand Aigle and Stephane Delauney. An aminoacylase activity from *Streptomyces ambofaciens* catalyzes the acylation of lysine on α -position and peptides on N-terminal position. *Engineering in Life Sciences*, (18/8):589-599, 2018
2. Boguslaw P Nocek, Danuta M Gillner, Yao Fan, Richard C Holz and Andrzej Joachimiak. Structural basis for catalysis by the mono- and dimetalated forms of the *dapE*-encoded N-succinyl-L,L-diaminopimelic acid desuccinylase. *Journal of Molecular Biology*, (397):617-626, 2010
3. JB Bonanno, Y Patskosky, J Freeman, KT Bain, S Hu, S Ozyurt, S Wassermann, JM Sauder, FM Raushel, SK Burley and SC Almo. Crystal structure of an uncharacterized metal-dependent hydrolase from *Pyrococcus furiosus*. *PDB X-ray structure validation report*, 2021

Inferring and comparing metabolism accross heterogeneous sets of annotated genomes using AuCoMe

Arnaud BELCOUR¹, Jeanne GOT¹, Méziane AITE¹, Ludovic DELAGE², Jonas COLLEN², Clémence FRIOUX³, Catherine LEBLANC², Simon DITTAMI², Samuel BLANQUART¹, Gabriel MARKOV^{2*} and Anne SIEGEL¹

¹Univ Rennes 1, Inria, CNRS, Irisa, 35000 Rennes, France

²CNRS - Sorbonne Université - Integrative Biology of Marine Models (UMR8227) - Station Biologique de Roscoff, Place Georges Teissier, 29680, Roscoff, France

³Inria, INRAE, Université de Bordeaux, France

Corresponding Authors: arnaud.belcour@protonmail.com, anne.siegel@irisa.fr

* Presenting

Comparative analysis of Genome-Scale Metabolic Networks (GSMNs) may yield important information on the biology, evolution, and adaptation of species [1]. However, it is impeded by the high heterogeneity of the quality and completeness of structural and functional genome annotations, which may bias the results of such comparisons [2]. To address this issue, we developed AuCoMe – a pipeline to automatically reconstruct homogeneous GSMNs from a heterogeneous set of annotated genomes without discarding available manual annotations [3]. We tested AuCoMe with three datasets, one bacterial, one fungal, and one algal, and demonstrated that it successfully reduces technical biases while capturing the metabolic specificities of each organism [4]. Our results also point out shared and diverging metabolic traits among evolutionarily distant algae, underlining the potential of AuCoMe to accelerate the broad exploration of metabolic evolution across the tree of life.

Acknowledgements

We acknowledge the GenOuest bioinformatics core facility <https://www.genouest.org> for providing the computing infrastructure. We also thank Erwan Corre (ABiMS Platform) and Pauline Hamon-Giraud for fruitful discussions. This work benefited from the support of the French Government via the National Research Agency investment expenditure program IDEALG (ANR-10-BTBR-04) and from Région Bretagne via the grant SAD 2016 - METALG (9673).

References

- [1] Changdai Gu, Gi Bae Kim, Won Jun Kim, Hyun Uk Kim, Sang Yup Lee. Current status and applications of genome-scale metabolic models. *Genome Biology* 20:121, 2019.
- [2] Delphine Nègre, Méziane Aite, Arnaud Belcour, Clémence Frioux, Loraine Brillet-Guéguen, Xi Liu, Philippe Bordron, Olivier Godfroy, Agneszka P. Lipinska, Catherine Leblanc, Anne Siegel, Simon M. Dittami, Erwan Corre, Gabriel V. Markov. Genome-Scale Metabolic Networks Shed Light on the Carotenoid Biosynthesis Pathway in the Brown Algae *Saccharina japonica* and *Cladosiphon okamuranus*. *Antioxidants* 8, 564, 2019.
- [3] <https://github.com/AuReMe/aucome>
- [4] Arnaud Belcour, Jeanne Got, Méziane Aite, Ludovic Delage, Jonas Collén, Clémence Frioux, Catherine Leblanc, Simon M. Dittami, Samuel Blanquart, Gabriel V. Markov, Anne Siegel. AuCoMe: inferring and comparing metabolisms across heterogeneous sets of annotated genomes. *Genome Research*, in press, 2023. Preprint: doi:10.1101/2022.06.14.496215.

Multiplex Network Exploration to define The landscape of Premature Aging Diseases

Cécile BEUST¹, Alberto VALDEOLIVAS^{1,2}, Anthony BAPTISTA^{1,3,4}, Galadriel BRIÈRE⁵, Ozan OZISIK¹ and Anaïs BAUDOT^{1,6}

¹ Aix Marseille Univ, INSERM, MMG, 13385, Marseille, France

² Roche Pharma Research and Early Development, Basel, Switzerland

³ School of Mathematical Sciences, Queen Mary University of London, London, E1 4NS, United Kingdom

⁴ The Alan Turing Institute, The British Library, London, NW1 2DB, United Kingdom

⁵ Aix Marseille Univ, CNRS, I2M, Marseille, France

⁶ Barcelona Supercomputing Center (BSC), 08034, Barcelona, Spain

Corresponding author: cecile.beust@univ-amu.fr, cecile.beust@etudiant.univ-rennes.fr

Premature Aging (PA) diseases form a group of rare genetic disorders mimicking certain aspects of physiological aging at an early age [1]. These diseases, usually monogenic, are clinically and genetically heterogeneous [1]. For example, the Hutchinson-Gilford Progeria syndrome can be caused by mutations in the *LMNA* or *ZMPSTE24* genes. These mutations cause important defects in nuclear lamina with a large range of cellular consequences and phenotypes including lipodystrophy, atherosclerosis and heart failure [1]. Another PA disease is the Classical Ehlers-Danlos syndrome, which is caused by mutations in several collagen genes (e.g., *COL1A1*, *COL5A1*, *COL5A2*), leading to hyperelastic skin and joint hypermobility phenotypes. Hence, in these two diseases, the PA-associated phenotypes seem to emerge from the perturbation of different cellular processes. However, to the best of our knowledge, a systematic description of the biological processes altered in PA diseases does not exist.

Biological interactions between genes and proteins offer the opportunity to study cellular processes on a large-scale. The diversity of physical and functional interactions can be represented as layers of a multiplex network, which can then be explored by appropriate algorithms. In particular, community identification algorithms allow extracting disease-associated subnetwork communities. As communities correspond to cellular processes, this approach allows identifying the processes likely perturbed in diseases.

In this work, we systematically identify the network communities associated with PA diseases. We first identified 68 genetic diseases annotated with a PA phenotype in the Human Phenotype Ontology (<https://hpo.jax.org/app/>). These diseases are caused by a total of 134 genes fetched from ORPHANET (<https://www.orpha.net/consor/cgi-bin/index.php>). We then built a multiplex biological network composed of 5 layers of interaction data, such as protein-protein or co-expression interactions. We then systematically extracted the network communities of the 68 PA diseases thanks to an iterative random walk with restart (it-RWR) implemented with the MultiXrank Python package [2]. We compared the 68 disease communities by computing their overlap with a Jaccard index, revealing that the 68 disease communities aggregate into 6 large clusters. Enrichment analyses of these clusters revealed different cellular processes, including for instance vesicle-mediated signaling, DNA repair and cell cycle, or mitochondrial processes. Overall, our approach revealed the molecular landscape of PA diseases and provides an improved understanding of the molecular mechanisms behind these disorders. Moreover, we compared the cellular processes of our landscape of PA with the hallmarks of aging described by Lopez-Otin and al. [3] and established links with physiological aging.

Acknowledgements

We thank the MarMaRa institute for funding this project.

References

- [1] Claire L. Navarro, Pierre Cau, and Nicolas Lévy. Molecular bases of progeroid syndromes. *Human Molecular Genetics*, 15(suppl.2):R151–R161, October 2006.
- [2] Anthony Baptista, Aitor Gonzalez, and Anaïs Baudot. Universal multilayer network exploration by random walk with restart. *Communications Physics*, 5(1):170, July 2022.
- [3] Carlos López-Otín, Maria A. Blasco, Linda Partridge, Manuel Serrano, and Guido Kroemer. Hallmarks of aging: An expanding universe. *Cell*, 186(2):243–278, January 2023.

PainterPipe: a pipeline for genetic variant fine-mapping using functional annotations

Zoe Gerber^{1,2}, Michel Fisun³, Hugues Aschard^{3,4} and Sarah Djebali¹

¹ IRSD, Université de Toulouse, INSERM, INRAE, ENVT, Université Toulouse III - Paul Sabatier (UPS), Toulouse, France

² Bordeaux Bioinformatics Master, Université de Bordeaux, Talence, France

³ Pasteur Institute, Université Paris Cité, Department of Computational Biology, Paris, France

⁴ Program in Genetic Epidemiology and Statistical Genetics, Harvard T.H. Chan School of Public Health, Boston, MA, United States

Corresponding Author: sarah.djebali@inserm.fr

Genome Wide Association Studies (GWAS) have identified thousands of genetic variants associated with common diseases. However, pinpointing variants that are truly causal remains a challenge. Indeed, GWAS results likely include a mix of causal variants and variants in linkage disequilibrium (LD, i.e. highly correlated) with the causal variants. In order to identify actual causal variants, fine-mapping methods have been developed. These methods use GWAS results and LD information, to assign to each variant a probability of being causal. In this field, PAINTOR [1] has become a standard, and one of its advantages is its ability to take into account functional annotations. Since a PAINTOR run requires a lot of pre- and post-processing steps, we decided to wrap all these steps into a Nextflow pipeline called PainterPipe (<https://github.com/sdjebali/PainterPipe>). PainterPipe uses three independent sources of information: GWAS summary statistics, LD information and functional annotations, to rank the variants according to their susceptibility to be involved in the development of the disease. The PAINTOR program is used to calculate the posterior probability of each SNP to be causal (a.k.a Bayesian fine-mapping). The resulting credible sets of variants are annotated with their biological functions and visualized using PAINTOR's visualization tool called CANVIS. This pipeline is implemented in the Nextflow pipeline specific language (DSL2) [2], can be run locally or on a slurm cluster and handles containerisation using Singularity [3]. It is designed to be modular and customizable, allowing for an easy integration of diverse functional annotations. To validate PainterPipe, we ran it (version 1.1.1) on GWAS results from the latest Coronary Artery Disease (CAD) meta-analysis involving 122,733 cases and 424,528 controls [4], and identified 149 loci with fine-mapping information. Out of the 161 CAD loci previously identified by the meta-analysis, and based on the presence of the lead SNP in our fine-mapped loci, 128 (78%) were common with our results. We additionally tested the impact of several input parameters of the pipeline, including the types of annotations used, on the pipeline's results on CAD, and also tested the pipeline on type 2 diabetes.

References

- [1] Kichaev G, Yang WY, Lindstrom S, Hormozdiari F, Eskin E, et al. (2014) Integrating Functional Data to Prioritize Causal Variants in Statistical Fine-Mapping Studies. *PLOS Genetics* 10(10): e1004722.
- [2] Di Tommaso, P., Chatzou, M., Floden, E. W., Barja, P. P., Palumbo, E., & Notredame, C. (2017). Nextflow enables reproducible computational workflows. *Nature Biotechnology*, 35(4), 316–319. doi:10.1038/nbt.3820
- [3] Kurtzer GM, Sochat V, Bauer MW (2017) Singularity: Scientific containers for mobility of compute. *PLOS ONE* 12(5): e0177459.
- [4] van der Harst P, Verweij N. Identification of 64 Novel Genetic Loci Provides an Expanded View on the Genetic Architecture of Coronary Artery Disease. *Circ Res.* 2018 Feb 2;122(3):433-443. doi: 10.1161/CIRCRESAHA.117.312086. Epub 2017 Dec 6. PMID: 29212778; PMCID: PMC5805277.

Exploitation des métadonnées et données génomiques d'agents pathogènes dans l'outil Bac'PACK pour améliorer les approches One HEALTH

Meryl VILA NOVA¹, Deborah MERDA¹ and Virginie CHESNAIS¹

¹ Laboratoire de Sécurité des Aliments - SPAAD, 14 rue Pierre Et Marie Curie, 94700, Maisons-Alfort, France

Le Centers for Disease Control and Prevention (CDC) définit le concept One Health comme une approche transdisciplinaire et multisectorielle incluant l'étude de la santé humaine, environnementale et animale ainsi que la sécurité des aliments, et qui a pour objectif principal l'étude et la prévention des maladies infectieuses émergentes. Mettre en œuvre ce concept nécessite le développement d'outils permettant de centraliser les systèmes de surveillance. Au sein de l'ANSES les activités des Laboratoires de Sécurité des Aliments (LSAI) et de Santé Animale (LSAn) portent sur l'étude et la surveillance de différents pathogènes impliqués dans les toxi-infections alimentaires : Salmonella, Listeria, Staphylococcus, Bacillus..., et sur des pathogènes responsables de maladies animales majeures du troupeau : Fièvre aphteuse, Tuberculose, Brucellose... L'essor des technologies haut-débit entraîne la création annuelle de plus de 4To de données utiles pour les activités de génomiques (Séquençage haut-débit, génotypage haut-débit...) des deux Laboratoires. Associées à ces résultats génomiques, les équipes du LSAI et LSAn doivent gérer de nombreuses métadonnées épidémiologiques et contextuelles d'origine variable. Afin de faciliter la gestion de ces nombreuses données, le service transversal de bio-informatique, le SPAAD, a créé l'outil Bac'PACK qui a pour objectif (i) la centralisation de données interopérable entre les différents domaines d'activité (épidémiologie et génomiques) grâce notamment au déploiement de référentiels internationaux pour la description des échantillons et du contexte des prélèvements ; (ii) la mise à disposition de ces données au travers des tableau de bord permettant la visualisation des données et leur requête et (iii) le partage d'informations vers les grandes instances sanitaires européennes et internationales. Ainsi, notre outil se compose de plusieurs bases de données MongoDB permettant de regrouper les informations disponibles pour chaque pathogènes étudiés dans les laboratoires ainsi qu'une interface graphique développée avec la librairie Streamlit facilitant l'enregistrement et l'interrogation de ces données. Au final, l'outil Bac'PACK que nous avons développé permet de suivre en temps réel, l'incidence d'un ensemble d'agents pathogènes sur les territoires nationaux et européens, et de détecter rapidement si ces derniers ont subi une modification

References

1. [Streamlit • The fastest way to build and share data apps.](#)
2. [MongoDB: La Plateforme De Données D'application | MongoDB](#)

The flies' route of 3GC^R *E. coli* dissemination in beef cattle farm: from ecosystem to molecular scale

Alann Caderhoussin¹, Gaëlle Gruel¹, David Couvin¹, Isaure Quétel¹, Matthieu Pot¹ Rémy Arquet², Alexis Dereeper¹, Philippe Glaser⁴, Jean-Christophe Bambou³, Antoine Talarmin¹, Séverine Ferdinand^{1*}

¹Institut Pasteur de la Guadeloupe, TRed-Path Unit, Les Abymes, Guadeloupe, France

²INRAE, Plateforme Tropicale d'Expérimentation sur l'Animal, Petit-Bourg, Guadeloupe, France

³INRAE, ASSET, Petit-Bourg, Guadeloupe, France

⁴Institut Pasteur, Ecology and Evolution of Antibiotic Resistance Unit, Paris, France.

INTRODUCTION

Antimicrobial resistance (AMR) is currently one of the most important public health problems in the world¹. It has dramatically increased morbidity and mortality in both humans and animals, with serious implications for the future treatment of human infections and for animal health and productivity². The emergence of AMR is mainly due to the selective pressure of antibiotics used in both human and veterinary medicine³. The use of antibiotic in food-producing animals raises concerns about the potential emergence and spread of resistant bacteria⁴.

Food-producing animals are potential reservoir of AMR, which can be transmitted to human pathogen *via* the food chain⁵, direct contact with animals, or release of manure into the environment⁶. The risk of zoonotic transfer from farm animals to humans in close contact with these animals is still largely unknown, but some studies have suggested a transfer of extended-spectrum β -lactamase (ESBL) *E. coli* or ESBLs genes from poultry or pigs to farm workers⁷. However, in most African surveys, among ESBLs, specific *bla*_{CTX-M-15} harboring clones are mainly identified in humans. In addition to this direct zoonotic transmission, other routes such as transmission *via* contaminated drinking water⁸, insects or a contaminated environment⁹, may be a risk factor for human colonization or infection.

Enterobacteriaceae carrying ESBLs are a worldwide clinical problem. ESBLs are mainly plasmid-encoded enzymes that confer extended resistance to β -lactam antimicrobials, due to their ability to inactivate cephalosporins. ESBL-producing bacteria are known as nosocomial pathogens or community-acquired pathogens, with *E. coli* being the most common. The occurrence of ESBL *E. coli* bacteria has also been widely reported as pathogens and/or colonizers in livestock, companion animals, zoo animals, and wildlife¹⁰ causing mastitis in dairy cattle, but most commonly livestock are asymptomatic carriers of ESBL producers¹¹.

The extent to which food-producing animals contributes to potential transmission of ESBL producers to humans is not well established. Though the direct transmission of AMR between animals and related environment, and human is still vague and debatable, the risk should not be neglected. By the way, we investigated AMR in healthy food-producing animals in Guadeloupe¹¹. We observed a moderate rate of ESBL-*E. coli* in a context of rational use of antimicrobials. Nevertheless, a hotspot of ESBL-*E. coli* *bla*_{CTX-M-15} carriers was discovered in beef cattle from farms free of third generation cephalosporins use. The factors driving AMR in food-producing animals requires meticulous investigation in order to introduce efficient initiatives to reduce resistance occurrence. These factors will be studied in environments reflecting real-life practices. The main objective was to assess the origin of ESBL-*E. coli* in cattle from an environment without 3GC-use. As we aimed to investigate AMR at the farm ecosystem level, we tested the hypothesis that ESBL-*E. coli* occurrence may originate from selective pressures other than 3GC use. Because oxytetracycline was widely used on farms for digestive pathology and to explain the occurrence of ESBL *E. coli* without the use of 3GC¹¹, we investigated oxytetracycline as a source of selective pressure. We also analyzed antiparasitic drugs widely used for tick-borne diseases or gastrointestinal strongle control, heavy metals as potential source of soil/plant contamination or antioxidants found in supplemental cattle feed. Because *bla*_{CTX-M-15} is of human origin, we also tested the hypothesis that ESBL-*E. coli* occurrence in cattle may originate from imported human isolates.

MATERIAL AND METHODS

Sampling and collection

We focused our investigations on the farms representing the highest rate of 3GCR *E. coli*¹¹. Between January 2018 and May 2019, beef cattle and goats, their environment, and food were screened for *E. coli*. Fresh fecal samples from cattle feces living in stabulation or in field, and goat feces living in stabulation were randomly collected just after excretion. Flying or resting flies around cattle feces, manure or goat breastfeeding food, and adult mosquitoes (*Culex* sp.) in unused goat feeders were trapped using a 6-volt mechanical aspirator. Drinking water and untreated water used for agriculture were sampled. Wastewater samples were collected downstream of the administration building. Cattle fodder, goat breastfeeding solubilized milk, milk powder, and pellets were collected aseptically. All samples were stored and transported in sterile cups or bags on ice to the Pasteur Institute laboratory within 4 h. Samples were stored at 4 °C and processed within 8 h of sampling.

Molecular identification of flies

Taxonomic assignment of fly species was performed by cytochrome oxidase I (COI) encoding gene PCR screening and Genbank sequence comparison. DNA was extracted individually from 7 morphologically distinct flies using the NucleoSpin® Tissue DNA Extraction Kit (Macherey-Nagel, Hoerd, France) according to the manufacturer's instructions. A fragment of the COI encoding genes (710 bp) was amplified in all flies by simplex PCR using the primers LCO1490 (forward) (5'-GGTCAACAAATCATAAAGATATTG G-3') and HCO2198 (reverse) (5'-TAAACTTCAGGGTGACCAAAAAATCA-3'). Amplified PCR products were sequenced (Eurofins, Köln, Germany) and compared to known COI gene sequences in the GenBank database, by multiple-sequence alignment using BLAST (<http://blast.ncbi.nlm.nih.gov/Blast.cgi>) for species assignment. All corresponding sequences were submitted to the phylogenetic tree reconstruction pipelines available on the Phylogeny.fr platform¹². The tree was constructed using the "Advanced" option, which allows statistical evaluation of branch support values using 100 bootstraps, and plotted using iTOL¹³.

Core genome comparative analysis

To assess the genomic relatedness and dynamics of ESBL *E. coli* transmission in a One Health strategy, high-throughput whole-genome sequencing was performed on 70 isolates at the "Plateforme de microbiologie mutualisée" (P2M) of the Pasteur International Bioresources Network (Pasteur Institute, Paris, France). DNA was extracted using a DNA Mini Kit (Qiagen). Libraries were prepared using the Nextera XT kit (Illumina), and sequencing was performed on the NextSeq 500 system (Illumina). Reads were trimmed and filtered using AlienTrimmer¹⁴. Genomes were assembled using SPAdes software¹⁵, and final quality was assessed using QUAST¹⁶. Maximum likelihood phylogenetic reconstruction was performed using RAxML software v.8.0.0¹⁷ with 100 bootstrap replicates on the "Université des Antilles" (UA) computing cluster. The GTR-CAT model was used for the phylogenetic reconstruction based on nucleotide sequences. The Slurm job scheduling system (<https://slurm.schedmd.com>) was used to parallelize the RAxML jobs on 16 cores and reduce the overall execution time. The tree was plotted using iTOL¹³. Where applicable, the multi locus sequence type (MLST) was identified based on the corresponding MLST scheme for *E. coli*¹⁸.

Multiplex long read sequencing and plasmid annotation

MinION ligation libraries were constructed from 1 µg unfragmented bacterial gDNA. Single flow cell sequencing data from 24 barcoded *E. coli* isolates were run on the MinION for 48 h with max. 512 sequencing channels available. Base-calling of MinION raw signals was performed using the Guppy v.3.2.10 data processing toolkit executed externally on the UA computing cluster and downloaded as FAST5 files. FASTQ files were extracted and split by barcode using the Epi2Me v.3.3.0 workflow available online (<https://epi2me.nanoporetech.com>). *De novo* genome assembly was performed using a hybrid strategy on combined nanopore long reads and previously available Illumina short reads. Fully resolved assemblies were obtained using the Unicycler pipeline¹⁹ with default parameters, implementing SPAdes *de novo* assembly and several rounds of Pilon polishing, and visualized using Bandage²⁰. Quality control of the nanopore data was performed using QUAST¹⁶. Plasmids were annotated with

RAST and graphically aligned with BRIG 0.95²¹. Plasmidfinder²² was used to identify replicon plasmid types. The presence of antibiotic resistance genes was identified using ResFinder²³. Mobilization module characterization was based on MOB-Suite²⁴ and virulence genes were analyzed using the VFDB (<http://www.mgc.ac.cn/VFs/main.htm>) web tools. Gene flow characterization was performed using set-joining analysis unions and subsequent Venn diagram illustration was performed using the web-based tool InteractiVenn²⁵. Three biological datasets on virulence and resistance genes provided by *in silico* analysis of whole genome sequences from bovine, fly and human ESBL-producing isolates were considered.

Sequence data availability

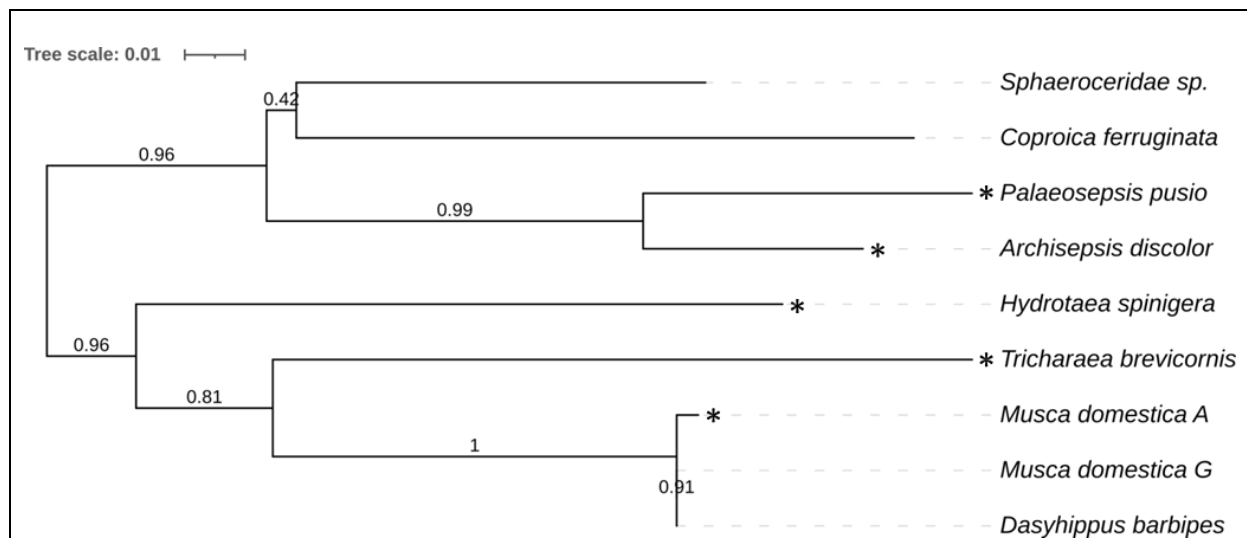
All the Illumina and Nanopore FASTQ data files generated and analyzed in the current study, will be deposited in the NCBI-SRA Public Archives.

RESULTS

ESBL *E. coli* carriage in Creole cattle and in a huge diversity of fly species

To assess the source of ESBL *E. coli*, the farm environment was investigated. A total of 41 samples were collected. At the farm ecosystem scale, ESBL *E. coli* were recovered from eight morphologically distinct flies collected around goats breastfeeding food and manure. ESBL *E. coli* were harbored by five distinct fly species. ESBL *E. coli* were sporadically isolated among the closely related species *Musca domestica* and *Dasyhippus barbipes* (Figure 1). ESBL *E. coli* were not detected in other environmental samples from the farm.

Figure 1. Genetic relatedness of flies based on cytochrome oxidase I gene fragment sequencing.

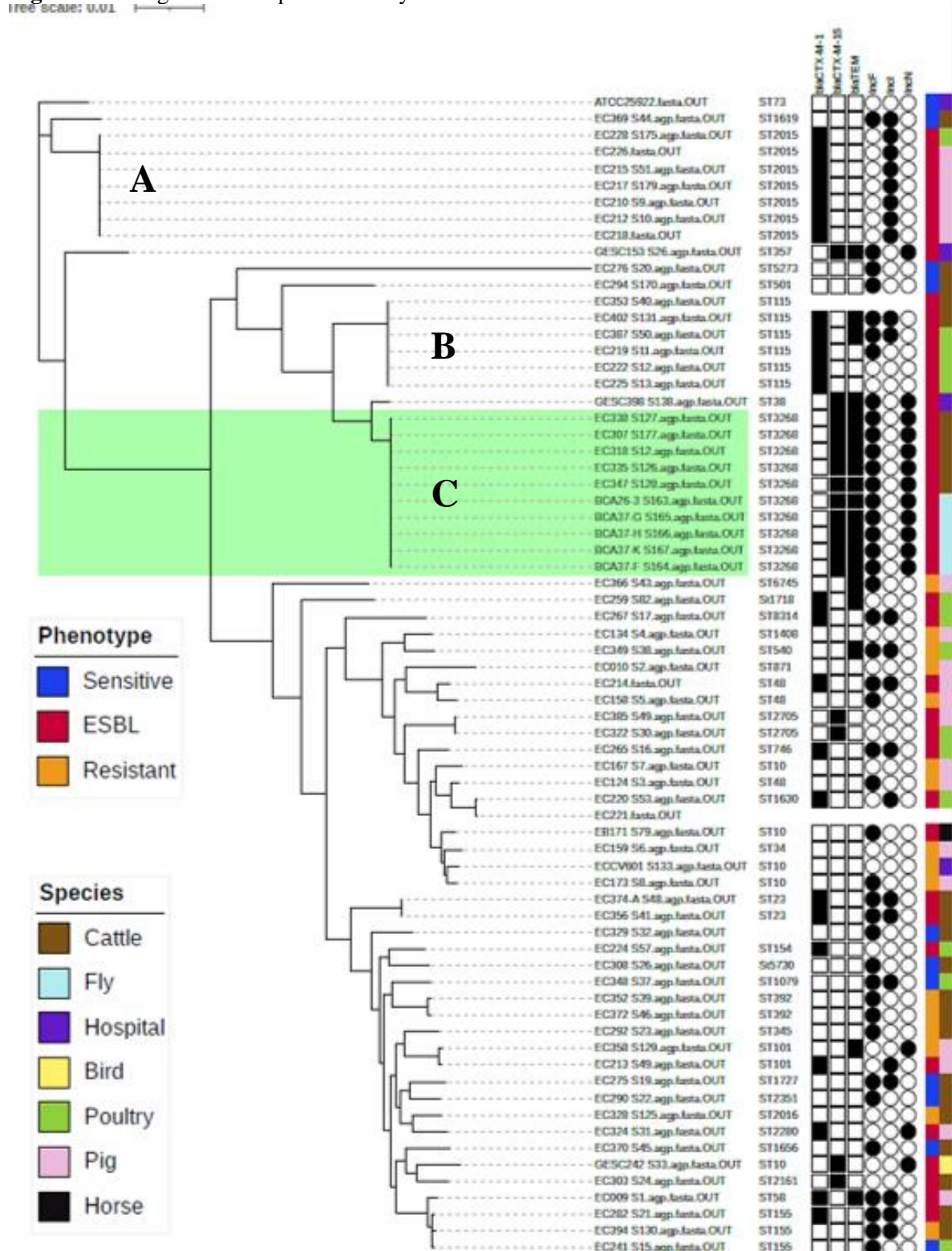


*Fly species carrying ESBL *E. coli* isolates were assigned by an asterisk.

An unexpected reservoir of flies and cattle harboring *bla*_{CTX-M-15} ESBL *E. coli*

At the bacterial level, the comparative core genome analysis of 70 *E. coli* revealed 47 distinct patterns (Fig. 2), with 29 (41.4%) isolates grouped into six clusters with similar core genome patterns comprising two to ten isolates, while 41 (58.6%) distinct patterns were not clustered. The three largest clusters (A, B, C), contained six to ten ESBL *E. coli*, harboring a *bla*_{CTX-M-1} (A, B) or a *bla*_{CTX-M-15} (C) gene. Cluster C showed a close genomic relationship between 10 fly and bovine ESBL *E. coli* ST3268 belonging to the ST38 complex and isolated from the same ecosystem. The ST38 complex isolates harbored plasmid-encoded *bla*_{CTX-M} genes of two different types (*bla*_{CTX-M-15} and *bla*_{TEM-1B}) that were inserted into two different plasmids.

Figure 2. Core genome comparative analysis of 70 ESBL *E. coli*



ST, sequenced type based on Oxford MLST classification.

Maximum likelihood phylogenetic tree of 70 ESBL *E. coli* isolates from Guadeloupe. Hosts and antimicrobial susceptibility phenotypes are indicated by vertical-colored stripes. Clusters are assigned by letters (A, B, C). A total of 40/70 isolates are extended-spectrum-b-lactamase (ESBL) producers. Corresponding resistance-coding genes, characterized by ResFinder, are indicated by black squares and plasmids by black circles.

An ecosystem with high potential for resistance spread and persistence

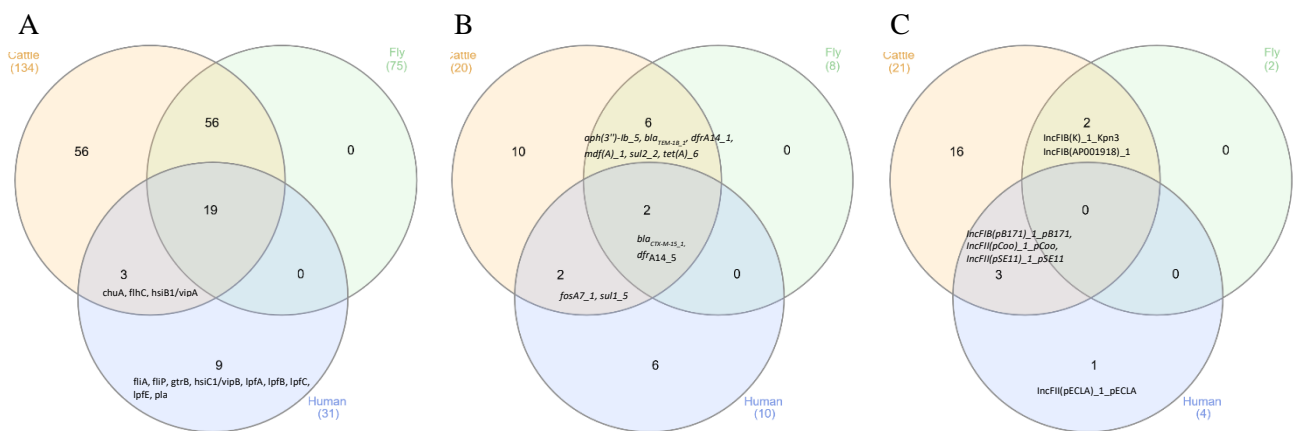
At the molecular level, plasmid sequencing allowed strains belonging to cluster C to be subdivided into two new subclusters (C.1 and C.2). Cluster C.1 included strains EC307, EC318, EC335, EC338, BCA37-F, -G, -H, -K, harboring three plasmid backbones from several incompatibility groups simultaneously and C.2 included two plasmids shared with C.1 (EC347, BCA26.3) (Figure 3). Unexpectedly, the *bla*_{CTX-M-15} gene was located in a transposon carried by a nonmobilizing replicative multireplicon plasmid IncFIB(K)_1_Kpn3-IncFIB(AP001918) cointegrates with a truncated IncN_1_AY046276 (248bp), which harbored many mobile genetic elements and several associated resistance genes. The conjugative replicon plasmid (IncN_1_AY046276, 50 979pb), absent in C2, carrying the *bla*_{TEM-1B} gene with a cassette of resistance gene and virulence genes involved in the type IV secretion (T4S) system was carried by ESBL *E. coli* strains common to cattle and flies. Mating experiments showed that the *bla*_{CTX-M-15} could be transferred by conjugation. The third, a phage-plasmid (47 973 pb), contained prophage regions from *Vibrio* and *Bacillus*, without resistance genes and a toxin HigB/antitoxin HigA system involved in pathogenicity regulation. ESBL *E. coli* ST38 complex-*bla*_{CTX-M-15/TEM-1B} was found in cattle, flies, and hospitalized patients (Figure 2). The human origin of a larger non-conjugative IncFIB(pB171)_1_pB171/*bla*_{CTX-M-15} plasmid (106 354 bp) was found in *Enterobacter* sp. collected in the wastewater of the administration building.



Cattle and flies as reservoirs for ESBL isolates enriched in pathogenicity factors

ESBL isolates from flies and cattle were enriched for virulence and resistance genes, respectively 78 ± 3 and 70 ± 14 on average per isolate vs 41 ± 0 for humans ($p=0.00001$), reflecting an animal reservoir of ESBL isolates carrying pathogenicity factors (virulence, resistance, replicon plasmid genes). To better characterize the circulation of virulence and resistance genes among ESBL-producing isolates collected from beef cattle ($n=26$), flies ($n=5$), and humans ($n=11$), were included in a Venn diagram recovered virulence genes ($n=143$), resistance genes ($n=26$), and replicon plasmid ($n=22$) provided by *in silico* whole genome sequencing data analysis (Figure 6). Similar genes and plasmids were shared between cattle and flies and between cattle and humans. However, replicon plasmids shared between cattle and flies ($n=2$) were different from those shared between cattle and humans ($n=3$). Indeed, two *bla*_{CTX-M-15} and *dfrA14* genes, common to the 3 groups, were carried by different plasmids depending on the biotope (Figure 6 B and C). In addition, no genes or plasmids were shared between flies and humans. This result highlights 2 distinct modes of transmission, depending on the type of host involved in the transmission pathway, and reveals a bottleneck in transmission between flies and humans. Since flies harbor the majority of pathogenicity factors that are never shared with humans, the fly reservoir of ESBL isolates and the main route of transmission of ESBL isolates to beef cattle by direct contact is easily identified. Since no gene was shared between flies and humans, there was no direct transmission. Furthermore, genes shared between flies and humans were associated with the bovine reservoir, suggesting (i) indirect transmission from flies to cattle and then from cattle to humans, and (ii) only to a minor extent by direct contact from cattle to humans. Our results suggest a transmission dynamic that reflects the spread of multiple persistent ESBL isolate lineages rather than a single epidemic circulating clone.

Figure 3. Pathogenicity factors and replicon plasmids distribution among biotopes



beta-lactamase-producing isolates from beef cattle, flies and humans.

DISCUSSION

This study of the origin of ESBL *E. coli*, in farm of food producing animals that are not exposed to third generation cephalosporins, allowed to identify a local cattle and flies reservoirs. ESBL *E. coli* from flies and cattle, ST3268 belonging to ST38 complex, were isolated in the same farm and were not found elsewhere in food-producing animals in Guadeloupe. The importance of ST38 among ESBL-producing *E. coli* has been described in human clinical isolates from Canada, Thailand, United Kingdom and Netherland, France, Germany, and Switzerland but also in livestock and food in Germany Mogolian wild bird and West African rats. As corroborate our results, ST38 has been described to be predominantly mediated by CTX-M-15²⁶. ST3268 has been recovered from humans, animals and the environment distributed worldwide. In our setting, the reservoir (flies and cattle) of ST3268 is delimited to one farm and the human compartment still seemed to be sporadically affected by this clone. Our results are consistent with a successful zoonotic *E. coli* ST38 lineage²⁷.

Nevertheless, plasmids are the main drivers of *bla*_{CTX-M} genes dissemination and the *bla*_{CTX-M-15} gene has been widely disseminated through various narrow host range plasmids replicon types in human including IncFIA, IncFII, IncHI2, IncM, and IncK²⁸. In our study, flies and cattle have accumulated multiple plasmids and genes and represent also a reservoir of resistant and virulent factors. The spread of *bla*_{CTX-M-15} in human isolates is a combination of successful clones and plasmids. In this way, dissemination of CTX-M-15 in human isolates is classically attributed to the expansion of the *E. coli* O25b-ST131 clone harbouring IncF plasmids. This strain habitually harbours the *bla*_{CTX-M-15} gene located on IncFII or FIA, FIB, and FII multireplicons. Indeed, IncF plasmids are low-copy-number plasmids, often carrying more than one replicon. It has been showed that *bla*_{CTX-M-15} can be carried by FIA-FIB, FIA-FIB-FII and FIB-FII multireplicons²⁹. In our setting, the *bla*_{CTX-M-15} gene was carried by a non-mobilizable multireplicon plasmid (2 IncFIB) cointegrating with a truncated IncN replicon. IncFIB-IncN multireplicon plasmids carrying the *bla*_{CTX-M-15} gene were also identified in *Salmonella* found in food sources in Colombia. One *E. coli* strain harbored an IncFIB(K) and IncFIB(AP001918)-type plasmid (ST38) was found in urban West African rats³⁰. Nevertheless, to the better of our knowledge, our multidrug resistance structure of IncFIB/*bla*_{CTX-M-15}/*bla*_{TEM1B} multi FIB replicon had never been described. IncN broad host-range replicon plasmid has been isolated in *Salmonella* spp. and *E. coli* from food producing animals and has been implicated in the dissemination and spread of CTX-M enzymes³¹. Indeed, homologous recombination between insertion sequences and transposons, appeared to be an important driver of new multiresistance regions³², and this may be coupled with recombination in plasmid backbones to reassort multiple IncF replicons cointegrates with IncN replicon as well as components of multiresistance regions. Even carried by a non-mobilizing IncF multireplicon

plasmid, the *bla*_{CTX-M-15} gene resides in environment of transposase coding genes and insertion sequences as markers of intracellular mobility, but also of other (*bla*_{TEM-1B} positive) plasmids and of many different Inc types (IncF, IncN conjugative) enabling inter-plasmidic transfer events and facilitating widespread dissemination of the *bla* genes between bacterial species and hosts. In addition, the IncN replicon contain a conjugative type IV secretion system (VirB1, virB3, virB5, virB8, virB10, virB11) usually described in IncF plasmids, mediating high-frequency transfer and accounts for the prevalence of these multi-drug resistance plasmids in environmental and clinical settings³³. It was observed that plasmids belonging to the IncN group are often colocalized with IncF plasmids³⁴. It has been proposed that in multireplicon plasmids, one replicon is strongly conserved due to the selective pressure imposed by the necessity of duplicating the plasmid, while the other is free to diverge³⁵.

The plasmid content of the ESBL isolates and the plasmid backbones indicate a significant plasticity of the genomes studied, probably reflecting an accumulation of rearrangements related to the selection pressures undergone. The *E. coli* ST3268 cluster deserves particular attention because it includes strains harboring several plasmids enriched in multidrug resistance and virulence factors, giving the strains a strong potential for sustainable spread. In a context where vectors such as flies favor the circulation of clones, the dynamics of transmission could be reactivated and a potential transmission from animals to humans is not inevitable.

REFERENCES

1. O'Neill, J. Antimicrobial Resistance : Tackling a crisis for the health and wealth of nations. *Review on Antimicrobial Resistance* <https://amr-review.org/sites/> (2014).
2. Eurosurveillance editorial team. WHO member states adopt global action plan on antimicrobial resistance. *Euro Surveill.* **20**, (2015).
3. Paterson, D. L. & Bonomo, R. A. Extended-Spectrum β -Lactamases: a Clinical Update. *Clin. Microbiol. Rev.* **18**, 657–686 (2005).
4. Landers, T. F., Cohen, B., Wittum, T. E. & Larson, E. L. A review of antibiotic use in food animals: perspective, policy, and potential. *Public Health Rep.* **127**, 4–22 (2012).
5. Marshall, B. M. & Levy, S. B. Food Animals and Antimicrobials: Impacts on Human Health. *Clin. Microbiol. Rev.* **24**, 718–733 (2011).
6. Heuer, H., Solehati, Q., Zimmerling, U., Kleineidam, K. & Schlöter Tanja Focks, Andreas Thiele Bruhn, Sören Smalla, Kornelia, M. M. Accumulation of sulfonamide resistance genes in arable soils due to repeated application of manure containing sulfadiazine. *Appl. Environ. Microbiol.* **77**, 2527–2530 (2011).
7. Dahms, C. *et al.* Occurrence of ESBL-Producing *Escherichia coli* in Livestock and Farm Workers in Mecklenburg-Western Pomerania, Germany. *PLoS One* **10**, 1–13 (2015).
8. Juhna, T. *et al.* Detection of *Escherichia coli* in biofilms from pipe samples and coupons in drinking water distribution networks. *Appl. Environ. Microbiol.* **73**, 7456–7464 (2007).
9. Dierikx, C. M., van der Goot, J. A., Smith, H. E., Kant, A. & Mevius, D. J. Presence of ESBL/AmpC -Producing *Escherichia coli* in the Broiler Production Pyramid: A Descriptive Study. *PLoS One* **8**, e79005 (2013).
10. Guyomard, S. *et al.* Antimicrobial resistance in wildlife in Guadeloupe (French West Indies): Dissemination of a single *bla*_{CTX-M-1/IncI1/ST3} plasmid scaffold in humans and wild animals. *Front. Microbiol.* In submission (2020).
11. Gruel, G. *et al.* Antimicrobial use and resistance in *Escherichia coli* from healthy food-producing animals in Guadeloupe. *BMC Vet. Res.* **17**, 116 (2021).
12. Dereeper, A. *et al.* Phylogeny.fr: robust phylogenetic analysis for the non-specialist. *Nucleic Acids Res.* **36**, W465-9 (2008).
13. Letunic, I. & Bork, P. Interactive tree of life (iTOL) v3: an online tool for the display and annotation of phylogenetic and other trees. *Nucleic Acids Res.* **44**, W242-5 (2016).
14. Criscuolo, A. & Brisse, S. AlienTrimmer removes adapter oligonucleotides with high sensitivity in short-insert paired-end reads. Commentary on Turner (2014) Assessment of insert sizes and adapter content in FASTQ data from NexteraXT libraries. *Front Genet* **5**, 130 (2014).
15. Bankevich, A. *et al.* SPAdes: a new genome assembly algorithm and its applications to single-

- cell sequencing. *J Comput Biol* **19**, 455–477 (2012).
16. Gurevich, A., Saveliev, V., Vyahhi, N. & Tesler, G. QUILT: quality assessment tool for genome assemblies. *Bioinformatics* **29**, 1072–1075 (2013).
 17. Stamatakis, A. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics* **30**, 1312–1313 (2014).
 18. Jolley, K. A., Bray, J. E. & Maiden, M. C. J. Open-access bacterial population genomics: BIGSdb software, the PubMLST.org website and their applications. *Wellcome open Res.* **3**, 124 (2018).
 19. Wick, R. R., Judd, L. M., Gorrie, C. L. & Holt, K. E. Completing bacterial genome assemblies with multiplex MinION sequencing. *Microb. Genomics* **3**, (2017).
 20. Wick, R. R., Schultz, M. B., Zobel, J. & Holt, K. E. Bandage: interactive visualization of de novo genome assemblies. *Bioinformatics* **31**, 3350–3352 (2015).
 21. Alikhan, N.-F., Petty, N. K., Ben Zakour, N. L. & Beatson, S. A. BLAST Ring Image Generator (BRIG): simple prokaryote genome comparisons. *BMC Genomics* **12**, 402 (2011).
 22. Carattoli, A. & Hasman, H. PlasmidFinder and In Silico pMLST: Identification and Typing of Plasmid Replicons in Whole-Genome Sequencing (WGS). *Methods Mol. Biol.* **2075**, 285–294 (2020).
 23. Zankari, E. *et al.* Identification of acquired antimicrobial resistance genes. *J Antimicrob Chemother* **67**, 2640–2644 (2012).
 24. Robertson, J. & Nash, J. H. E. MOB-suite: software tools for clustering, reconstruction and typing of plasmids from draft assemblies. *Microb. genomics* **4**, (2018).
 25. Heberle, H., Meirelles, G. V., da Silva, F. R., Telles, G. P. & Minghim, R. InteractiVenn: a web-based tool for the analysis of sets through Venn diagrams. *BMC Bioinformatics* **16**, 169 (2015).
 26. Alghoribi, M. F. *et al.* Antibiotic-resistant ST38, ST131 and ST405 strains are the leading uropathogenic Escherichia coli clones in Riyadh, Saudi Arabia. *J. Antimicrob. Chemother.* **70**, 2757–2762 (2015).
 27. Fonseca, E. L., Morgado, S. M., Caldart, R. V & Vicente, A. C. Global Genomic Epidemiology of Escherichia coli (ExPEC) ST38 Lineage Revealed a Virulome Associated with Human Infections. *Microorganisms* **10**, (2022).
 28. Coque, T. M., Baquero, F. & Canton, R. Increasing prevalence of ESBL-producing Enterobacteriaceae in Europe. *Euro Surveill. Bull. Eur. sur les Mal. Transm. = Eur. Commun. Dis. Bull.* **13**, (2008).
 29. Marcadé, G. *et al.* Replicon typing of plasmids in Escherichia coli producing extended-spectrum beta-lactamases. *J. Antimicrob. Chemother.* **63**, 67–71 (2009).
 30. Schaufler, K. *et al.* Clinically Relevant ESBL-Producing K. pneumoniae ST307 and E. coli ST38 in an Urban West African Rat Population. *Front. Microbiol.* **9**, (2018).
 31. Carattoli, A. Plasmids in Gram negatives: molecular typing of resistance plasmids. *Int. J. Med. Microbiol.* **301**, 654–658 (2011).
 32. Partridge, S. R., Zong, Z. & Iredell, J. R. Recombination in IS26 and Tn2 in the evolution of multiresistance regions carrying blaCTX-M-15 on conjugative IncF plasmids from Escherichia coli. *Antimicrob. Agents Chemother.* **55**, 4971–4978 (2011).
 33. Liu, X., Khara, P., Baker, M. L., Christie, P. J. & Hu, B. Structure of a type IV secretion system core complex encoded by multi-drug resistance F plasmids. *Nat. Commun.* **13**, 379 (2022).
 34. Szmolka, A. *et al.* First report on IncN plasmid-mediated quinolone resistance gene qnrS1 in porcine Escherichia coli in Europe. *Microb. Drug Resist.* **17**, 567–573 (2011).
 35. Sýkora, P. Macroevolution of plasmids: A model for plasmid speciation. *J. Theor. Biol.* **159**, 53–65 (1992).

Poster 1

Alteration of transposable element mutation rate by nucleosome positioning

Fabien SASSOLAS¹, Jeremy BARBIER², Jean-Nicolas VOLFF¹ and Benjamin AUDIT²

¹ Institut de Génomique Fonctionnelle de Lyon, UMR5242, Ecole Normale Supérieure de Lyon, Centre National de la Recherche Scientifique, Université Claude Bernard-Lyon 1.

² Ens de Lyon, CNRS, Laboratoire de Physique, F-69342 Lyon, France.

Corresponding Author: fabien.sassolas@ens-lyon.fr

A physical model of nucleosome formation showed that, for one third of the human genome, the positions of nucleosomes are directly encoded in the DNA sequence through statistical positioning of 2-3 nucleosomes at the border of regions predicted to be too energetically costly to bend around histones, called “nucleosomal inhibitory energy barriers” or NIEBs. In human, 1.5 million NIEBs have been detected, one every 1.5-1.8 kbp without any particular association to genomic regions but a depletion at gene TSSs. We found NIEBs in all higher eukaryote genomes analyzed, with similar characteristics in terms of length and distribution among genomes.

GC content oscillates coherently with nucleosome positioning at the NIEB borders, NIEBs and linker regions presenting a lower GC content than sequences covered by nucleosomes. By analyzing mutations since the last common ancestor between *Mus musculus* and *Mus caroli*, we found that mutation patterns in unique sequences explain this oscillation, with mutations increasing the GC content being larger at nucleosome positions compared to NIEBs and linkers while the opposite is observed for mutations increasing the AT content, as previously observed in mutation profiles in human since divergence from chimpanzee [1].

Alu are ~300 bp primate specific SINE transposable elements. Alu are composed of two GC rich monomers separated by an A-rich linker and terminated a poly-A tail. They thus present a GC content oscillation similar to the one observed at NIEB borders. Interestingly, more than half of the Alu elements are located with their poly-A at a NIEB border having their two GC-rich arms over the nucleosome positions [2, 3]. However, the pattern of mutation over these Alu does not match the one obtained for unique sequences, that could be due to restricted mutation contexts. For instance, there is a maxima of weak (A/T) → strong (G/C) mutations at the position of the linker in the Alu sequences whereas it's a minima in the unique sequences. Nevertheless when controlling for the mutation context, we observed that in Alu poly-A tails, the ATA → AAA mutations are enriched compared to unique sequences in the NIEB borders in contrast with mutations breaking poly-As (AAA → ABA) that are similar in both cases. This result suggests some Alu specific mutational patterns at NIEB borders that require further investigation.

We are now extending our analysis to other transposable elements as well as determining mutations in other species (e.g. in fish). We notably focus on B1 elements, a *Mus* specific SINE, that we found to follow the same positioning pattern at NIEB borders as Alu elements. This will allow us to address the phylogenetic extend of the specific sequence evolution at NIEB borders described in human.

References

- [1] Guénola Drillon, Benjamin Audit, Françoise Argoul, et Alain Arneodo. « Evidence of Selection for an Accessible Nucleosomal Array in Human ». *BMC Genomics* 17, n° 1: 1-20, 2016.
- [2] Frédéric G. Brunet, Benjamin Audit, Guénola Drillon, Françoise Argoul, Jean-Nicolas Volff, et Alain Arneodo. « Evidence for DNA Sequence Encoding of an Accessible Nucleosomal Array across Vertebrates ». *Biophysical Journal* 114, n° 10: 2308-16, 2018.
- [3] Barbier, Jérémy, Fabien Sassolas, Cédric Vaillant, Jean-Nicolas Volff, Frédéric G. Brunet, et Benjamin Audit. « Insertion of Alu elements impacts sequence-mediated nucleosome positioning ». In *JOBIM: 23ièmes Journées Ouvertes en Biologie, Informatique et Mathématiques*.

Challenges of genomic data generation for non-model complex species

Guillaume DORE¹, Frédérique BARLOY-HUBLER² and Dominique BARLOY¹

¹ UMR DECOD (Ecosystem Dynamics and Sustainability) Institut Agro, IFREMER, INRAE, 65 Rue de Saint-Brieuc, 35042, Rennes, France

² UMR 6553 ECOBIO, CNRS, Université de Rennes 1, 263 Avenue du Général Leclerc, 35042, Rennes, France

Corresponding Author: guillaume.dore@agrocampus-ouest.fr

Ludwigia grandiflora subsp. *hexapetala* (*Lgh*) is an invasive aquatic plant very common in France [1]. *Lgh* has the ability to colonize wet meadows [2]. To understand the (epi)genetic mechanisms underlying this ability to change environment, genomic data are necessary. *Lgh* is a non-model species for which there are no genomic resources available. This species is decaploid ($2n=10x=80$) [3] and its genome is very big (1,4 Gb). To generate genomic resources, we have chosen to first assemble organellar genomes. The *Lgh* plastome was recently assembled as two haplotypes [4]. Obtaining the mitochondrial genome proved to be more complicated as mitogenomes are highly variable in sequence and size [5]; numerous repeated sequences are present and possible chloroplastic gene insertions [6]. Moreover, plant mitogenomes are poorly represented in database compared to plastomes. In this study we present the different strategies used to assemble and obtain *Lgh* mitogenome.

DNA long fragments were extracted from *Lgh* buds then were sequenced using two types of technology: Oxford Nanopore (long reads) and Illumina Mi-seq (short reads). We decided to use a hybrid methodology to combine benefits of both sequencing reads to assemble *Lgh* mitogenome as it was efficient to assemble *Lgh* plastome [4]. As we didn't have any reference genome, three strategies were used to build the mitochondrial genome: the first strategy concerned an *a priori* approach using available mitogenomes of the Myrtales order species as references; the second strategy used conserved mitochondrial sequences from close species coding DNA sequences (CDS = sequence of protein-coding genes) [5] and the third strategy concerned *de novo* assembly. These three strategies were compared and combined when necessary.

The first strategy did not permit to get *Lgh* mitogenome, probably due to phylogenetic distance with available Myrtales order species mitogenomes and high variability in intergenic regions. The second approach generated 39 CDS consensus sequences, efficiently used to select mitochondrial reads. Finally, by combining second and third strategies, we were able to assemble the complete *Lgh* mitogenome which contains two circular molecules M1 and M2 of respectively 544,782 bp and 166,796 bp. We identified repeated sequences and chloroplastic insertions in the mitogenome.

We offer an efficient hybrid strategy to assemble *de novo* mitogenome in a non-model species. This assembly will serve for further transcriptomic analysis (RNA-seq). Due to the complexity and size of *Lgh* nuclear genome and because we are interested in functional analysis, we have rather chosen to assemble a reference transcriptome instead of a genome.

References

- [1] S. Dandelot. Invasive *Ludwigia* spp. of southern France: History, Taxonomy, Biology and Ecology. Thesis, 2004.
- [2] J. Haury, F. Noël, M. Bozec, J. Coudreuse, J. Guil, et al.. Importance of *Ludwigia grandiflora* as invasive weed on meadows and pastures in Western France. 3rd International Symposium on Weeds and Invasive Plants, 2011, Ascona (CH), Switzerland.
- [3] L. Thouvenot, J. Haury and G. Thiebaut. A success story: water primroses, aquatic plant pests. Aquatic conservation: Marine and Freshwater Ecosystems, 2013.
- [4] F. Barloy-Hubler, A-L. Le Gac, C. Boury, E. Guichoux, D. Barloy. The existence of two haplotypes chloroplast genomes in *Ludwigia* species: *de novo* assembly combining long- and short-reads. BMC Plant Biology, 2023 (submitted).
- [5] A. J. Alverson, X. X. Wei, D. W. Rice, D. B. Stern, K. Barry and J. D. Palmer. Insights into the Evolution of Mitochondrial Genome Size from Complete Sequences of *Citrullus lanatus* and *Cucurbita pepo* (Cucurbitaceae). Molecular Biology and Evolution, 2010.
- [6] J. P. Mower, D. B. Sloan and A. J. Alverson. Plant Mitochondrial Genome Diversity: The Genomics Revolution. Plant Genome Diversity Volume 1, 2012.

Poster: Characterization of Transposable Elements in Pangenomes

Somia SAIDI¹, Johann CONFAIS² and Hadi QUESNEVILLE²

¹ Université Paris-Saclay, INRAE, URGI, 78026, Versailles, France

Corresponding Author: somia.saidi@inrae.fr

Transposable elements (TEs) are mobile DNA elements that can invade genomes by transposition. Despite their reputation as parasitic sequences, they can enrich the genomes with functional novelties that foster genome evolution.

The impact of TEs in a genome can be explored by searching for their insertions. Individuals of the same species independently undergo TE insertions causing inter-individual genetic variability. This variability between individuals is the basis of the natural selection that leads to an increased adaptation of individuals to their environment.

A way to search for the potential role of TEs in host adaptation is through a pangenomic approach. The TE pangenome can be described by (i) TE insertions present in all individuals of the species (core-genome), (ii) insertions present only among a subset of individuals (dispensable-genome) or (iii) ecogenome when the individuals share the same environment, and finally (iv) individual-specific insertions.

Current pangenome analysis methods are based on the alignment of reads from different genomes of the species to an assembled reference genome. But, the advent of the third-generation sequencing makes now possible to better approach this question using several assembled genomes of the same species to avoid the bias introduced by a single reference genome.

I will present a new pipeline, called panREPET, which identifies TE copies in a pangenome from several assembled genomes. There is therefore no dependency on a reference genome. This pipeline identifies copies shared by a group of individuals.

This pipeline has been tested on 54 genomes of *Brachypodium distachyon* to describe its pangenomic compartments.

Evolution and molecular characterization of internal targeting signals in mitochondrial proteins

Clément GALAN¹ and Ingrid LAFONTAINE¹

¹ Institut de Biologie Physico-Chimique - UMR7141, 13 Rue Pierre et Marie Curie, 75005 Paris, France

Corresponding Author: ingrid.lafontaine@ibpc.fr

Abstract

At the origin of the emergence of eukaryotic lifestyle, the primary endosymbiosis event occurred. This was followed by a massive transfer of genetic material from the newly formed endosymbiont to the cell nucleus. Thus, the vast majority of proteins found in organelles are translated in the cytosol and imported in the organelle. In most cases, this is done via a transit peptide at the N-terminal end of the protein that allow its interaction with the organelle outer membrane followed by its translocation across the outer and inner membrane. Additional targeting-like sequences located in the internal part of the protein has been detected, that seem to play a role in the import mechanism, as they would improve the import of some protein precursors by binding to an outer membrane receptor [1].

We analyzed the distribution of the internal signals, coupled with the description of their physico-chemical properties among the whole yeast proteome. We also analyzed the conservation of these internal signals among gene families over the entire tree of life. Our preliminary results show that these signals are present only certain proteins and would have facilitated proper import of ancestral bacterial genes transferred to the nucleus.

References

1. S. Backes et al., « Tom70 enhances mitochondrial preprotein import efficiency by binding to internal targeting sequences », *Journal of Cell Biology*, vol. 217, n° 4, p. 1369-1382, janv. 2018, doi: 10.1083/jcb.201708044.

Evolution of corn metabolites detoxification in *Ostrinia*

Mariam EL KHATIB¹, Ana Paula Zotta-Mota², Anthony Bretaudeau³, Vincent Calcagno⁴ and Frédérique Hilliou¹

¹ Institute Sophia Agrobiotech ID team, 400 Rte. des Chappes, 06903, Sophia Antipolis, France

² Institute Sophia Agrobiotech GAME team, 400 Rte des Chappes, 06903, Sophia Antipolis, France

³ BIPAA, Plate-forme GenOuest IRISA-INRIA, Campus de Beaulieu 35042 Rennes cedex, France

⁴ Institut Sophia Agrobiotech M2P2 team, 400 Rte des Chappes, 06903, Sophia Antipolis, France

Corresponding Author: frederique.hilliou@inrae.fr

1. Introduction

With the introduction of Corn into Europe and Asia around 500 years ago, a host shift was observed in the two allopatric *Ostrinia* species *Ostrinia nubilalis* (ECB) and *Ostrinia furnicalis* (ACB) to corn. This shift occurred independently. These two species are also sympatric with the Eurasian *Ostrinia scapulalus* (ABB), which still favors the original dicot host[1]. This occurrence presents a unique opportunity to study convergent evolution and highlight putative independent adaptations in the three *Ostrinia* species. We are interested in investigating the repertoire of P450 genes in *Ostrinia*. P450 is a large superfamily of heme monooxygenase enzymes. They are involved in detoxification, making them candidates for the acquired ability in ECB and ACB to detoxify corn metabolites[2]. By manually annotating P450 genes, we aim to compare their repertoires and perform phylogenetic analysis between the three species, which could unravel the origin of this adaptation.

2. Annotation of P450 genes and phylogenetic analysis

To annotate P450 genes on the sequenced *Ostrinia* genomes, we are using “Web Apollo” [3], a tool to edit and locate genes based on evidence such as BLAST results of previously curated P450 genes from Lepidopteran species against our sequence of interest. We are also using HMMER [4], to detect P450 homologs in *Ostrinia* proteomes. We are using a second P450 hmm profile, based on a database of manually curated lepidopteran P450 sequences. After annotating P450 genes in the three *Ostrinia* species, we aim to perform phylogenetic analysis using RAxML-NG, a phylogenetic tree inference tool based on maximum likelihood[5]. Upon comparing the phylogeny of P450 genes in the three species and other lepidopteran corn pests, we will gain an insight into the presence of orthologous P450 genes, blooming and duplication events, or loss of P450. We will try to associate these events with potential adaptation to corn as a new host.

Acknowledgments

This work was supported by the “ANR Muscado” to Vincent Calcagno. We thank the BioInformatics Platform for Agroecosystem Arthropods (<https://bipaa.genouest.org/>) for hosting the *Ostrinia* genomes.

References

1. H. Alexandre *et al.*, “When History Repeats Itself: Exploring the Genetic Architecture of Host-Plant Adaptation in Two Closely Related Lepidopteran Species,” *PLoS One*, vol. 8, no. 7, p. e69211, Jul. 2013, doi: 10.1371/JOURNAL.PONE.0069211.
2. D. F. V. Lewis, *Guide to Cytochromes P450 Structure and Function*, 2nd ed, New York, Taylor & Francis, 2002, pp 1-18.
3. E. Lee *et al.*, “Web Apollo: A web-based genomic annotation editing platform,” *Genome Biol*, vol. 14, no. 8, pp. 1–13, Aug. 2013, doi: 10.1186/GB-2013-14-8-R93/TABLES/1.
4. S. R. Eddy, “HMMER User’s Guide Biological sequence analysis using profile hidden Markov models,” 2020, Accessed: May 02, 2023. [Online]. Available: <http://hmmer.org>
5. A. M. Kozlov, D. Darriba, T. Flouri, B. Morel, and A. Stamatakis, “RAxML-NG: a fast, scalable and user-friendly tool for maximum likelihood phylogenetic inference,” *Bioinformatics*, vol. 35, no. 21, pp. 4453–4455, Nov. 2019, doi: 10.1093/BIOINFORMATICS/BTZ305.

Evolution of habenular asymmetries in gnathostomes: a transcriptomic approach

Hélène Mayeur*^{†1}, Lucile Guichard , Léo Michel , Ronan Lagadec , and Sylvie Mazan[‡]

¹UMR 7232 – Observatoire océanologique de Banyuls – France

Résumé

Habenulae are bilateral epithalamic structures found in vertebrates, involved in the integration of sensory perceptions and the regulation of adaptive behaviors. A remarkable feature of these structures is that they display asymmetries between left and right in many vertebrates. Comparisons across gnathostomes against the catshark suggest that some asymmetries have an ancient origin in the taxon but that lineage-specific diversifications have occurred in osteichthyans, resulting in independent losses in tetrapods and neoteleosts. In order to gain insight into these diversifications and obtain an unbiased characterization of ancestral asymmetries in gnathostomes, we focused on three gnathostomes: a chondrichthyan, the catshark *Scyliorhinus canicula*, an actinopterygian, the Senegal bichir *Polypterus senegalus*, and a sarcopterygian, the lungfish *Protopterus annectens*. We thus conducted a transcriptomic comparison between left and right habenulae in each of these species. To this end, total RNA was extracted and Illumina cDNA libraries were constructed from left or right manually dissected habenulae explants from juveniles (three replicates per species). Sequencing yielded a total of 8500 million reads which were mapped on NCBI gene model references for each species. Statistical analyses led to lists of 2208, 607 and 4171 putative differentially expressed transcripts between left and right habenulae in *S. canicula*, *P. annectens* and *P. senegalus* respectively. Cross-species pairwise comparisons highlight a conserved core of genes differentially expressed both in the catshark and lungfish (89 genes), or catshark and bichir (88 genes), albeit not necessarily with the same laterality. They also reveal a set of 14 genes sharing asymmetric expressions in all three species, which may reflect ancestral asymmetries in gnathostomes. Experimental validations by *in situ* expression profilings in the species studied are ongoing. Together with ongoing functional analyses in the catshark, they will be crucial to gain insights into the mechanics and evolutionary trends which may underlie this pattern of conservations.

Mots-Clés: transcriptome, evolution, differential expression, comparison, phylogeny

*Intervenant

[†]Auteur correspondant: helene.mayeur@obs-banyuls.fr

[‡]Auteur correspondant: sylvie.mazan@obs-banyuls.fr

Genomics landscape shaped by transposable elements through rust pathogens history

Emma CORRE¹, Emmanuelle MORIN¹, Cécile LORRAIN², Sébastien DUPLESSIS¹

¹ Université de Lorraine, INRAE, UMR IAM, 54280 Champenoux, France

² Plant Pathology Group, Institute of Integrative Biology, ETH Zürich, Zürich, Switzerland

Corresponding Author: emma.corre@inrae.fr

Abstract

Rust fungi (Basidiomycota, Pucciniomycotina, Pucciniomycete, Pucciniales) represents the most expanded order of obligate biotrophic plant pathogens. These ubiquitous pathogens threaten food security, infecting hosts such as soybean, wheat or coffee. The genomes of rust fungi are remarkably larger (from 80Mb to 1.2Gb) than the average genome size of other Pucciniomycotina species (~40Mb). Rust fungi genomes also harbor higher gene and repeat content, including transposable elements (TEs). TEs are mobile genetic elements, which are well-known for their major impact on genome evolution and they likely explain genome expansion and evolution of rust fungi. However, we lack a systematic annotation of TEs in the genomes of rust fungi and in other Pucciniomycotina to understand how and when TEs affected their genome architecture.

To better understand the evolutionary history of rust fungi genomes, we selected 15 genomes of Pucciniales and 6 genomes of other species in the Pucciniomycotina. We plan to annotate and compare the *de novo* TE repertoires in the 21 selected species with the REPET pipeline. We will further assess inter- and intra-species TE insertion ages combining relative – using comparative genomics approach – and absolute dating – with Kimura's distance of LTR retrotransposons – of TE expansion events. We will finally explore the potential association between TE burst events and expansion of specific multi-gene families in Pucciniales using Computational Analysis of gene Family Evolution (CAFÉ) combined with functional annotation.

Preliminary results indicate that TE repertoires are diverse among 15 Pucciniomycotina genomes annotated so far among the 21 selected. Overall, we observed a higher proportion of retrotransposons in genomes of Pucciniales compared to other Pucciniomycotina. Notably, genomes of Pucciniales seem to be enriched in LINE retrotransposons with ~16 % of the total TE families on average in Pucciniales versus ~1.6 % in other Pucciniomycotina. Our preliminary analysis of LTR age insertion indicates that although a small proportion of ancient copies are detected in multiple punctual bursts between 2 to 20 Mya, most of the LTR inserted between 0 and 2 Mya in the genomes of Pucciniomycotina. This result suggests recent retrotransposon activity in the Pucciniomycotina. The number of recent LTR copies is drastically higher in the genomes of Pucciniales compared to those of other Pucciniomycotina suggesting that LTRs have recently contributed to the genome size expansion of Pucciniales.

Our first results suggest a switch in the genomic landscape of Pucciniales shaped by specific TE bursts in the recent history of Pucciniomycotina.

Highlighting the evolution towards antibiotic cross-resistance in *E. coli* biofilms exposed to biocides using large-scale comparative genomics

Pierre Lemée¹, Raphaël Charron^{1,2}, Marine Boulanger¹, Paméla Houée¹, Christophe Soumet¹, Romain Briandet², Arnaud Bridier¹

¹ Antibiotics, Biocides, Residues and Resistance Unit, Fougères Laboratory, Anses, 35300 Fougères, France

² Paris-Saclay University, INRAE, AgroParisTech, Micalis Institute, 78350, Jouy-en-Josas, France

Corresponding Author: pierre.lemee@anses.fr

The emergence of multi-drug resistant bacteria over the past decades is a major global health problem. Understanding the factors that are responsible for the selection and propagation of these bacteria has become essential. One possibility is the use of disinfectants containing biocides along the food chain for decontamination. Bacterial populations chronically exposed to biocides, sometimes at sublethal concentrations can evolve toward resistance to these molecules. Moreover, evidences demonstrated that adaptation to biocides could participate to decrease bacterial susceptibility against other families of antimicrobials as antibiotics through common mechanisms such as overexpression of efflux pumps for instance [1]. If an increasing number of studies indeed underlined the development of cross-resistance between biocide and antibiotics in planktonic populations including in pathogens [2], bacteria mostly gather in surface-associated communities called biofilms along the food chain and exhibit specific resistance and adaptive traits. The ANR BAoBAb project aims to study the mechanisms of biofilm adaptation to biocides and the role of this collective process in the spread of resistance to antibiotics.

In this study, we wanted to highlight the mechanisms of cross-resistance to antibiotics in biofilms exposed to biocides. To do so, we selected 10 *Escherichia coli* strains producing biofilms isolates from food chain and exposed them to two different biocides (benzalkonium chloride and triamine) during 5 weeks. Each week, strains were selected on selective agar for their resistance against antibiotic (rifampicin). Whole genome sequencing was then performed on 212 selected resistant strains. The genome structure of the 10 parental bacteria was compared to the biofilms of each bacterial lineage that became resistant to antibiotics to identify the metabolic pathways and genes impacted. Mutations in genes belonging to lipopolysaccharide biosynthesis pathway were mostly associated to triamine exposure whereas benzalkonium chloride exposure is rather associated to mutations in the ribose metabolic pathway. Together, such results suggest that selection of antibiotic cross-resistance likely occurred through specific biocide-dependant evolutive pathways.

References

1. Bridier, A. *et al.* Impact of cleaning and disinfection procedures on microbial ecology and Salmonella antimicrobial resistance in a pig slaughterhouse. *Scientific Reports* 9, (2019).
2. Guérin, A. *et al.* Exposure to Quaternary Ammonium Compounds Selects Resistance to Ciprofloxacin in *Listeria monocytogenes*. *Pathogens* 10, 220 (2021).

La phylodynamie pour le suivi épidémiologique de la fièvre catarrhale en Guyane

Déborah MERDA ¹, Mathilde GONDARD ², Emmanuel BREARD ², Yannick BLANCHARD ³, Corinne SAILLEAU ², Virginie CHESNAIS ¹, Stephan ZIENTARA ²

¹ University Paris Est, ANSES, SPAAD unit, Maisons-Alfort location, F-94701 Maisons-Alfort, France

² UMR 1161 ANSES/INRA/ENVA, Université Paris-Est ANSES Maisons-Alfort, Maisons-Alfort, France.

³ Unit of Viral Genetics and Biosafety, ANSES, Laboratory of Ploufragan, Ploufragan, France.

Corresponding Author: deborah.merda@anses.fr

Les analyses de phylodynamie permettent de mettre en évidence les mécanismes évolutifs des virus, et notamment les phénomènes de réassortiments génomiques pouvant être à l'origine de modifications de la pathogénicité des virus réassortis. La fièvre catarrhale ovine (FCO) est une maladie virale qui affecte les ruminants, dans le monde entier. Le génome de la FCOvirus est composé de 10 segments d'ARN, dont 2 codent les protéines de surface (capside externe), déterminant le sérotype. En Guyane, de nombreux sérotypes de ce virus circulent pour lesquels un suivi épidémiologique de 2010 à 2020 a été réalisé. Afin d'analyser les processus éventuels de réassortiments qui ont lieu dans cette région, une collection de 43 souches a été séquencée en utilisant la technologie IonTorrent. Les génomes ont été assemblés avec MIRA [1]. L'alignement des séquences correspondant aux 10 segments d'ARN du virus ont été alignés avec l'outil mafft [2]. Les analyses de diversité génétique réalisées avec le package R pegas [3] ont permis de mettre en évidence des niveaux de diversité variables entre les différents sérotypes, suggérant des phénomènes de réassortiments. Ces événements ont pu être mis en évidence en utilisant une approche d'inférence bayésienne avec l'outil BEAST2 [4] et le package CoalRe [5], dont le modèle de coalescent prend en compte les événements de réassortiment et a déjà été utilisé pour analyser l'histoire évolutive des virus Influenza. Le taux de réassortiment moyen par lignée et par année a pu être inféré, et est de 0.28. Ce taux est équivalent à celui inféré chez H1N1 [5]. En conclusion, le nombre de réassortiments est important chez ce virus, et implique 8 des 10 segments du virus (seules les 2 gènes codant pour les protéines de la capsid externe demeurent systématiquement associés). Ces réassortiments peuvent avoir lieu entre souches de sérotypes différents. Ceci représente un risque d'émergence de nouvelles souches dont la virulence ne serait pas connue.

1. Chevreaux, B., Pfisterer, T., Drescher, B., Driesel, A. J., Muller, W. E., Wetter, T., & Suhai, S. Using the miraEST assembler for reliable and automated mRNA transcript assembly and SNP detection in sequenced ESTs. *Genome Research*, (14), 1147–1159, 2004.
2. Katoh, Misawa, Kuma, Miyata. MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. *Nucleic Acids Res.* (30) 3059-3066, 2002.
3. Paradis Emmanuel. 2009. PEGAS: an R package for population genetics with an integrated-modular approach. *Bioinformatics Applications note*, 0, 1-2, 2009.
4. Bouckaert, R., Heled, J., Kühnert, D., Vaughan, T., Wu, C.H., Xie, D., Suchard M. A., Rambaut A., Drummond, A. J. Beast 2: A software platform for Bayesian evolutionary analysis. *PLoS Computational Biology*, 10, e1003537, 2014.

5. Müller, N. F., Stolz, U., Dudas, G., Stadler, T., & Vaughan, T. G.. Bayesian inference of reassortment networks reveals fitness benefits of reassortment in human influenza viruses. *Proceedings of the National Academy of Sciences*. 117, 17104-17111, 2020.

Testing pangenomic tools for structural variant detection in non-model organisms

Sukanya DENNI^{1,2}, Quynh TRANG-BUI³, Nicolas LAPALU⁴, Véronique DECROOCQ³, Christophe PLOMION² and Ludovic DUVAUX²

¹ Université de Rouen Normandie, 1 rue Thomas Becket, F-76821 Mont-Saint-Aignan cedex, France

² BIOGECO UMR 1202 INRAE, Univ. Bordeaux, 69 route d'Arcachon, CS80227, F-33610 Cestas, France

³ BFP, UMR 1332 INRAE, Univ. Bordeaux, 71 Avenue Edouard Bourlaux, CS20032, F-33883 Villenave d'Ornon, France

⁴ BIOGER UMR 1290 INRAE, Campus Agro Paris-Saclay - INRAE-AgroParisTech, 22 place de l'Agronomie, CS20040, F-91123 Palaiseau Cedex, France

Recent advances in the field of pangenomics - e.g. development of tools to build and manipulate pangenome graphs (e.g. Siren et al. 2021) - now allow to detect and genotype structural variants (SVs) at a population scale for a reasonable cost. However, most pangenomics tools were developed with the human genomes in mind and it is currently unclear to which extent they can be applied to non-model organisms with different genomic features (e.g. higher genetic diversity, different transposable element content, different ploidy, ancestral whole genome duplications) and less resolved genome assemblies. Here, we tested the robustness - to various genomic conditions deviating from the human model - of one of the two common steps performed in pangenomic analyses: SV genotyping from short reads using vg Giraffe.

Using three datasets of previously assembled genomes from species with heterogeneous levels of genome complexity and genetic diversity (a fungi, an apricot tree and a European white oak), we tested the capacity of vg Giraffe to robustly and accurately map reads for individuals with high divergence to the pangenome graph. By simulating reads with increasing mutation rates, we found that vg Giraffe is able to accurately map reads diverging by up to 3% from the graph and that mapping quality remains above 30 for divergence up to 2% confirming that this tool is suitable for species with high genetic diversity. Further analyses are under progress to assess the effect of more complex variations such as INDELS in SV calling quality.

In further investigations, we will evaluate pangenome graph building tools like minigraph and PGGB and their impacts on pangenome quality and SV genotyping. We will estimate the minimal depth of sequencing required for accurate SV genotyping using either approach and check the SV genotyping accuracy of vg Giraffe. The pipeline and tests used for these analyses are available on gitlab as Snakemake/Singularity workflows.

Jouni Sirén, et al., Pangenomics enables genotyping of known structural variants in 5202 diverse genomes. *Science* **374**, abg8871 (2021). DOI: [10.1126/science.abg8871](https://doi.org/10.1126/science.abg8871)

The complete mitochondrial genome of coffee leafminer *Leucoptera coffeella*, a major pest of coffee crops

Mateus PEREIRA DOS SANTOS^{1,2}, Ana PAULA ZOTTA MOTA¹, Roberto TOGAWA³, Maria APARECIDA CASTELLANI², Érika VALERIA SALIBA ALBUQUERQUE FREIRE³, Frédérique HILLIOU¹

¹ Université Côte D'Azur, INRAE, CNRS, ISA, 400 route des Chappes BP -167, F-06903 Sophia Antipolis, France

² State University of Southwestern Bahia/UESB, Estrada do Bem-Querer, km-04, 3293-3391, Vitória da Conquista - BA, Brasil

³ Embrapa Genetic Resources and Biotechnology, 70770-917, Brasília-DF, Brazil

Corresponding Author: mateus.pereira-dos-santos@inrae.fr frederique.hilliou@inrae.fr

Introduction

The coffee leafminer, *Leucoptera coffeella* is a key pest that causes economic losses to coffee crops that occurs in several countries (e.g., Mexico, and Brazil). Their damages to the plant consist of the formation of mines on the leaves, typical of leafminer insects. In larval phase, insects consumes the palisade parenchyma of coffee leaves which reduces photosynthetic area, resulting in defoliation and reducing productivity [1]. Problems such as resistance to insecticides have been reported for *L. coffeella* due to the wide pesticides use [2]. Therefore, it is important to find new ways to control this pest in coffee plantations, seeking sustainable strategies, such as biologics methods and integrated pest management programs. The absence of basic genomic information about this insect is a limitation for the development of population control strategies and monitoring of insect communities based on molecular strategies and also to gain in-depth knowledge about the evolutionary history of coffee leafminer. In this study, we presented the first complete mitochondrial genome of *L. coffeella*.

Assembly and annotation of mitochondrial genome

To assemble the mitochondrial genome of the leafminer, we retrieved the closest mitochondrial genome from NCBI (*L. malifoliella*). This sequence was used as seed for the software Aladin. The cleaned reads from the whole genome sequence were mapped to the seed and we retrieve one single contig (16,407 bp) corresponding to the mitochondrial genome of *L. coffeella*. The final mitochondrial sequence was annotated using MITOS2 web-server [3] with invertebrate's genetic code.

Results

We found a total of 37 genes, including 13 protein-coding genes (PCGs), 22 transfer RNA genes (tRNAs) and 2 ribosomal RNA genes (rRNAs). To study the phylogenetic relationship of the different *Leucoptera* species publicly available, we will present the result of phylogeny for each *L. coffeella* mitochondrial genes (13 PCGs and 2 rRNAs) using MAFFT v 7.475 [4] with 23 arthropods species as a phylogeny of concatenated mitochondrial genes. Our results will allow the development makers for *L. coffeella* identification in the field. It will also guide future studies on the evolutionary history of the coffee leafminer and its relationship with other lepidoptera's pests and insects. The analysis of the mitochondrial genome of *L. coffeella* will also contribute to expanding knowledge about its taxonomy. These results represent advances for the generation of new biotechnological tools for implementing the management of lepidopterans in coffee crops.

Acknowledgements

Consórcio Brasileiro de Pesquisa e Desenvolvimento do Café - CBP&D/Café and Conselho Nacional de Desenvolvimento Científico e Tecnológico - CNPq

References

1. J. Dantas, I. O. Motta, L. A. Vidal, E. F. M. B. Nascimento, J. Bilio, J. M. Pupe, A. Veiga, C. Carvalho, R. B. Lopes, T. L. Rocha, L.P. Silva, J. R. Pujol-Luz, E.V. S. Albuquerque. A comprehensive review of the coffee leaf miner *Leucoptera coffeella* (Lepidoptera: Lyonetiidae) —a major pest for the coffee crop in Brazil and others neotropical countries. *Insects*, (12/12): 1130, 2021.
2. S.A. Leite, M.P. dos Santos, D. R. da Costa, A.A. Moreira, R.N.C. Guedes, M.A. Castellani. Time-concentration interplay in insecticide resistance among populations of the Neotropical coffee leaf miner, *Leucoptera coffeella*. *Agricultural and Forest Entomology*, (23/2): 232-241, 2021.
3. Alexander Donath, Frank Jühling, Marwa Al-Arab, Stephan H Bernhart, Franziska Reinhardt, Peter F Stadler, Martin Middendorf, Matthias Bernt. Improved annotation of protein-coding genes boundaries in metazoan mitochondrial genomes. *Nucleic Acids Research*, (47/20): 10543–10552, 2019.
4. Kazutaka Katoh, John Rozewicki, Kazunori D Yamada. MAFFT online service: multiple sequence alignment, interactive sequence choice and visualization. *Briefings in Bioinformatics*, (20/4): 1160–1166, 2019.

Free-living forms and EVEs of brown algae viruses

Karine Massau^{*†1,2}, Erwan Corre^{‡1}, Loraine Brillet-Guéguen^{§1}, Alexandre Cormier^{¶3},
and Mark J Cock^{||1}

¹Station biologique de Roscoff – Sorbonne Université, Centre National de la Recherche Scientifique –
France

²Université de Rouen Normandie – Normandie Université – France

³Institut Français de Recherche pour l'Exploitation de la MER – Institut Français de Recherche pour
l'Exploitation de la MER - IFREMER, Institut Français de Recherche pour l'Exploitation de la Mer
(IFREMER) – France

Résumé

Brown algae are a phylogenetic group that evolved independently of red and green algae¹. Viruses infect this group of organisms and are in fact commonly present in the genomes of brown algae as inserted proviruses or as endogenous viral elements (EVEs)². The brown algal viruses (phaeoviruses) are classified as nucleocytoplasmic long DNA viruses (NCLDVs) because they have large double-stranded DNA genomes. We know that free-living forms of these viruses exist, but it is extremely rare to observe them. We therefore wanted to see if we could find traces of free forms of the virus in raw algal genome assembly data (akin to metagenomic data), corresponding to 41 brown algae samples.

Different approaches were evaluated. Firstly, we used the Virify analysis pipeline³ which allowed us to identify several candidate viral sequences within the different samples. However, further analysis of these candidates suggested misclassification of viral sequences as bacteriophage sequences. We therefore used an alternative approach to identify potential viral sequences based on aligning of raw assembly contigs to three viral reference genomes available in public databases, and then attempted to align these potential viral sequences to the algal reference genome, to remove potential inserted viral forms (proviruses or EVEs). Potential viral sequences that did not (or poorly) match the algal genome were retained. Then, gene prediction and functional annotation were performed on these sequences in order to identify clusters of viral genes. However, very few potential viral sequences were found in addition to the sequences that had already been identified as EVEs in the algal genomes, and very few of the new sequences were of sufficient size to predict enough genes to find a potential sequence belonging to a free form of the virus. This result may be due in part to the very small number of phaeoviruses described in the reference databases and to difficulty assigning hitherto unknown or phylogenetically very distant functions to the known sequences in these databases. We are therefore continuing our work by now analyzing the EVEs (Endogenous Viral Elements) embedded in algal genomes, focusing initially on the genomic insertion sites of the EVEs.

*Intervenant

†Auteur correspondant: kmassau@sb-roscoff.fr

‡Auteur correspondant: erwan.corre@sb-roscoff.fr

§Auteur correspondant: loraine.gueguen@sb-roscoff.fr

¶Auteur correspondant: alexandre.cormier@ifremer.fr

||Auteur correspondant: cock@sb-roscoff.fr

Mots-Clés: Comparative genomic, Virus, Brown algae

Insect invaders analysis pipeline - What happens to the genetic load when alien species become invasive?

Barbara Porro^{*†1}, Mathieu Zallio^{*1}, Emeline Deleury¹, Aurélie Blin¹, Mathieu Gautier², Eric Lombaert¹, and Martine Da Rocha¹

¹INRAE - ISA – Institut national de recherche pour l’agriculture, l’alimentation et l’environnement (INRAE), Université Côte d’Azur – France

²Centre de Biologie pour la Gestion des Populations – Centre de Coopération Internationale en Recherche Agronomique pour le Développement, Institut de Recherche pour le Développement, Institut National de Recherche pour l’Agriculture, l’Alimentation et l’Environnement, Institut Agro Montpellier, Université de Montpellier – France

Résumé

Biological invasions, and more specifically those enhanced by anthropogenic activities, are becoming more and more frequent. They in turn have a huge impact on these activities, but also on biodiversity. During biological invasions, we generally observe a decrease in genetic diversity in the introduced populations compared to the native populations. This is basically explained by a reduced number of individuals at the origin of the settlement of the new population. To explain the success of an invasion despite low genetic diversity, one hypothesis proposes a purging of the genetic loading (*i.e.*, a loss of deleterious mutations) in the introduced populations (1,2).

To study the evolution of the genetic load, we performed whole-genome pool-sequencing of native and invasive populations of a dozen insect species (3). This large amount of genetic data and species diversity will provide one of the broadest overviews of what happens during a biological invasion, but will also require the development of a specific pipeline. Therefore, using the workflow manager Nextflow and Singularity containers, we developed a pipeline to analyze this specific poolseq data.

References

1. Glémin Sylvain. 2003. "How are deleterious mutations purged? Drift versus nonrandom mating." *Evolution* 57:2678–2687, <https://doi.org/10.1111/j.0014-3820.2003.tb01512.x>
2. Daly, Ella Z., Olivier Chabrerie, Francois Massol, Benoit Facon, Manon C. M. Hess, Aurélie Tasiemski, Frédéric Grandjean, Matthieu Chauvat, Frédérique Viard, Estelle Forey, Laurent Folcher, Elise Buisson, Thomas Boivin, Sylvie Baltora-Rosset, Romain Ulmer, Patricia Gibert, Gabrielle Thiébaud, Jelena H. Pantel, Tina Heger, David M. Richardson, and David Renault. 2023. "A Synthesis of Biological Invasion Hypotheses Associated with the Introduction–Naturalisation–Invasion Continuum." *Oikos* 1–29, <https://doi.org/10.1111/oik.09645>
3. Evolution of the genetic load during biological invasions. <https://anr.fr/Projet-ANR-19-CE02-0010>

*Intervenant

†Auteur correspondant: barbara.porro@inrae.fr

Mots-Clés: Population genomics, Pool, sequencing, Singularity Containers, Nextflow

phylEntropy, a web-based tool for various data visualization applications

Damien Cazenave¹, Davy Regalade¹, Vincent Moco¹, Erick Stattner², Wilfried Segretier², Jimmy Nagau²,
Isaure Quétel¹, Sebastien Breurec¹, Severine Ferdinand¹, Alexis Dereeper¹, David Couvin¹

¹ Transmission, réservoir et diversité des pathogènes, Institut Pasteur de la Guadeloupe,
Les Abymes, Guadeloupe, France

² Laboratoire de Mathématiques Informatique et Applications (LAMIA), Université des
Antilles, F-97154, Pointe-à-Pitre, Guadeloupe, France

Corresponding Author: dcazenave@pasteur-guadeloupe.fr

Abstract

Background:

Data visualization is an area of data science that uses graphical tools to translate large amounts of data into understandable visuals. Several libraries, tools and modules exist to draw charts from data. The development of visualization tools is necessary in specific and broader data analysis processes. These developments could also be of interest in leveraging and simplifying the interpretation/rendering of sequencing and other biological data.

Results:

We developed phylEntropy, a web-based data analysis and visualization application, allowing us to perform data analysis and visualization based on various criteria (quantitative and qualitative variables). The tool brings together several analysis modules representing different research themes: genomics, phylogeny/clustering, biostatistics, measures of biological diversity, multivariate analyses, and machine learning. Users can upload a semicolon separated CSV file to the website and get various kinds of outputs (graphs, trees, charts, tables, and other visualizations/statistics). The first column of our file contains the identifier (ID) of each “object”. The following columns contain quantitative variables (one number per column). The last two columns (representing qualitative variables) generally contain character strings (text).

The web application is available at: <https://github.com/dcazenav/PhylEntropy>

Conclusions:

With phylEntropy, we aim at providing the scientific community with a platform offering simple data analysis and visualizations in a user-friendly manner. This web application also helps users in Machine Learning predictions in function of given input data.

Acknowledgements

We thank the researchers and doctoral students who provided new datasets to help us improve the phylEntropy web application.

References

1. Pedregosa F, Varoquaux, Gaël, Gramfort A, Michel V, Thirion B, Grisel O, et al. Scikit-learn: Machine learning in Python. *Journal of machine learning research*. 2011;12 (Oct):2825–30.

Poster 2

Benchmarking read mapping on pangenomic variation graphs

Hajar BOUAMOUT¹, Benjamin LINARD¹ and Matthias ZYTNICKI¹

¹ Unité de Mathématiques et Informatique Appliquées, INRAE, Castanet-Tolosan, France

Corresponding Author: benjamin.linard@inrae.fr

1 Pangenomes and variation graphs

A pangenome represents the total genetic diversity of a species or a species complex. One can describe pangenomes in terms of gene presence / absence variations (PAVs) but a more recent alternative aims to integrate full length genomes in a sequence graph [1]. In particular, pairwise alignments of a genome set can be used to build a “Variation Graph” (VG) in which nodes represent words of genome fragments and edges represent the contiguity of the words in one to several genomes (e.g., edges are associated to genome subsets). Each input genome consequently corresponds to a particular path in the graph.

It has been showed that VGs can improve variant calling and genotyping processes [2]. Indeed, variant calling is often based on the mapping of linear query sequences to linear reference genomes, thus biasing the prediction to regions present in the reference and limiting the identification of structural variations that are specific to other genomes. In particular, biases are reduced when large structural variations (>50bp) are targeted.

2 Sequence to graph mapping

Identifying new variants via a pangenome graph requires a compulsory preliminary step of querying sequence to graph mapping. Several approaches have been proposed (see [3] for a review), with algorithms dedicated to either long or short sequence reads. Most of these tools use a 2 steps method with: a) the identification of candidate sub-graphs showing similarity to the query read via different techniques of indexation followed by b) a refined alignment in the selected sub-graphs (generally via a technique of partial order alignment).

In practice, it remains unclear how this preliminary will impact further variants predictions. In particular, it has yet to be shown how resilient will be the different approaches to diverse mutation and indel rates.

3 Tools evaluation

The poster will present the results produced by Hajar Bouamout during her Master internship dedicated to the evaluation of sequence to graph read mapping tools. It will briefly describe the main ideas behind the algorithms proposed by 4 tools: GraphAligner [4], vg map [5], vg giraffe [5] and Minichain [6], and report benchmarks made with these tools.

References

- [1] Paten B, Novak AM, Eizenga JM, Garrison E. *Genome graphs and the evolution of genome inference*. Genome Res. 2017 May;27(5):665-676.
- [2] Garrison E, Sirén J, Novak AM, Hickey G, Eizenga JM, Dawson ET, Jones W, Garg S, Markello C, Lin MF, Paten B, Durbin R. *Variation graph toolkit improves read mapping by representing genetic variation in the reference*. Nat Biotechnol. 2018 Oct;36(9):875-879.
- [3] Shuo Wang, Yong-Qing Qian, Ru-Peng Zhao, Ling-Ling Chen, Jia-Ming Song, Graph-based pan-genomes: increased opportunities in plant genomics, Journal of Experimental Botany, Volume 74, Issue 1, 1 January 2023, Pages 24–39
- [4] Rautiainen M, Marschall T. *GraphAligner: rapid and versatile sequence-to-graph alignment*. Genome Biol. 2020 Sep 24;21(1):253.
- [5] Hickey G, Heller D, Monlong J, Sibbesen JA, Sirén J, Eizenga J, Dawson ET, Garrison E, Novak AM, Paten B. *Genotyping structural variants in pangenome graphs using the vg toolkit*. Genome Biol. 2020 Feb 12;21(1):35.
- [6] Ghanshyam Chandra, Chirag Jain. *Sequence to graph alignment using gap-sensitive co-linear chaining*. bioRxiv 2022.08.29.505691;

Development of a supervised ctDNA analysis pipeline to improve minimal residual disease monitoring in lymphoma

Lucie Gomes*^{†1,2,3}, Pierre-Julien Vially³, Vincent Sater³, Marie-Hélène Delfau-Larue⁴, Elodie Bohers³, Mathieu Viennot³, Philippe Ruminy³, and Fabrice Jardin³

¹French Connect plateforme – Institut Carnot Calym – France

²Master Bioinformatique, Modélisation et Statistique – Université Rouen Normandie – France

³INSERM U1245 – Centre de Lutte Contre le Cancer Henri Becquerel Normandie Rouen, Université de Rouen Normandie, Institut National de la Santé et de la Recherche Médicale - INSERM, Université de Rouen Normandie – France

⁴INSERM U955 – Assistance Publique-Hôpitaux de Paris (AP-HP), Hôpital Henri Mondor – France

Résumé

In onco-hematology, the analysis of cfDNA extracted from the blood gives access to the patients' tumoral mutational profile throughout the course of the disease. However, in order to detect and anticipate patient resistance and relapse, it is necessary to develop algorithms to improve the detection performance of low-frequency variants and to discriminate them from sequencing errors. We present here a new supervised pipeline developed by the national bioinformatics platform of the Carnot Institute CALYM, which uses the technology of UMIs and phased variants to enable minimal residual disease monitoring in lymphoma.

Mots-Clés: liquid biopsy, ctDNA, NGS sequencing, variant calling, phased variant, minimal residual disease, lymphoma

*Intervenant

[†]Auteur correspondant: c.luciegomes@gmail.com

P-GRe : un nouveau pipeline automatique pour l'annotation précise des pseudogènes

Sébastien CABANAC¹ Catherine MATHE-DEHAIS¹ et Christophe DUNAND¹

¹ Laboratoire de Recherches en Sciences Végétales, 24 Chemin de Borde Rouge, 31326, Castanet-Tolosan, France

Corresponding Author: sebastien.cabanac@univ-tlse3.fr

Les pseudogènes sont des séquences dérivées de gènes, parfois issus de la duplication d'un gène dit « parent » ou de la rétrotranscription d'un ARN messager [1], ayant perdu leur fonction suite à une ou plusieurs mutations génétiques ou à l'insertion de la (rétro)copie à un locus dépourvu de séquence promotrice. Parfois considérés comme du « junk DNA » car non-traduits, la capacité de certains pseudogènes à être transcrits a pourtant été démontré et était soupçonnée depuis plus de 40 ans [2]. Aujourd'hui, le rôle de la transcription des pseudogènes comme régulateur de l'expression des gènes a été mis en évidence à de nombreuses reprises [3]. Malgré cela, il n'existe à ce jour aucun pipeline bioinformatique entièrement automatique permettant à la fois la prédiction des positions des pseudogènes sur un génome, leurs structures en pseudo-exons, pseudo-introns, pseudo-CDS, *etc*, la construction de leurs séquences génomiques et pseudo-codantes, et une approximation des protéines « virtuellement codées » par ceux-ci. Dans ce poster, nous présentons P-GRe, un pipeline entièrement automatique et ne nécessitant que la séquence du génome à annoter et les coordonnées des gènes (et leurs structures) sur ce génome.

P-GRe est un pipeline en cours de développement capable d'annoter les pseudogènes ainsi que leurs structures sur un génome à annoter. Actuellement testé sur le génome de la plante modèle *Arabidopsis thaliana*, P-GRe permet de retrouver entre 83% et 85% des pseudogènes connus chez cette espèce, contre 81,3%, 79,8% et 6,0% (*sic*) pour trois logiciels couramment utilisés pour la prédiction de pseudogènes (respectivement PseudoPipe, Shiu's pipeline et PSF [4]). Même si nous n'avons pas encore de données chiffrées, P-GRe semble également trouver plus précisément la structure des pseudogènes, et par conséquent produire des protéines « virtuelles » de meilleure qualité (*i.e.* qui s'alignent mieux avec des protéines réellement synthétisées). Les étapes du pipeline sont détaillées ci-après, en insistant sur les points les plus « innovants » (4, 5 et 6) qui ont permis d'obtenir ces résultats.

1. Les coordonnées des gènes sont utilisées pour masquer leurs séquences sur le génome à annoter. 2. Le protéome est généré au format FASTA à l'aide des coordonnées des gènes, de leurs structures et du génome. 3. Les séquences en acides aminés sont blastées contre le génome. 4. Les résultats obtenus sont filtrés. Au lieu d'appliquer des filtres avec un seuil strict sur la longueur et le pourcentage d'identité des hits obtenus, un filtre sur l'identité est appliqué seulement, mais un seuil est calculé pour chaque hit en fonction de sa longueur, avec pour logique que « les hits les plus courts sont acceptés, mais doivent avoir un pourcentage d'identité plus fort ». 5. Les hits qui se chevauchent et dont les longueurs combinées ne sont pas divisibles par trois sont considérés comme des traces de décalage du cadre de lecture. L'ensemble des séquences protéiques possibles en décalant le cadre de lecture pour chaque position où les hits se chevauchent est alors construit, et la séquence protéique qui s'aligne le mieux avec la protéine codée par le gène parent du pseudogène est utilisée pour trouver la position précise du décalage du cadre de lecture. 6. Pour repérer les introns, une première protéine est construite et est alignée avec la protéine codée par le gène parent par l'algorithme de Gotha, affiné par un « *a priori* Lindley-like process » développé spécialement pour P-GRe, puis par une recherche des sites d'épissage. 7. Un codon d'initiation et un codon stop sont recherchés aux alentours de la protéine si nécessaire.

References

1. Yusuf Tutar. Pseudogenes. *Comparative and functional genomics*, 2012
2. David V Goeddel *et al.* The structure of eight distinct cloned human leukocyte interferon cDNAs. *Nature* (290):20–26, 1981
3. Ryan C Pink *et al.* Pseudogenes: Pseudo-functional or key regulators in health and disease? *RNA* (17/5):792–798, 2011
4. Jin Xiao *et al.* Pseudogenes and Their Genome-Wide Prediction in Plants. *International journal of molecular science* (17/12), 2016

Visual Representation of Genomes

Erick STATTNER¹, Wilfried SEGRETIER¹, Nalin RASTOGI² and David COUVIN²

¹ Laboratory of Mathematics, Computer Science and Applications (LAMIA), University of the French West Indies, France

² WHO Supranational TB Reference Laboratory, Tuberculosis and Mycobacteria Unit, Institut Pasteur de la Guadeloupe, France

Corresponding author: erick.stattner@univ-antilles.fr

1 Introduction

The evolution of genome sequencing has seen great technical advances [1], which have led to a significant cost reduction and an increase in sequencing speed and accuracy. However, the complexity of genomes is a major challenge in the field of genomic analysis. Several approaches in the literature have been proposed in order to address this problem through the transformation of the whole genome into new representation spaces [2]. In this work, we present a new model, called *G2P* (*from Genome to Pixels*), which aims to transform a genome into a two-dimensional image. The intuition behind this approach is that the genetic variations, as well as the areas of interest in the genome, can be modeled through different levels of contrast in an image. The approach has been developed into a JAVA a tool (<https://github.com/estattner/G2P-Model>) and can be applied on any genome.

2 Methodology

The model we propose performs with three parameters: (i) the genome GN , (ii) the portion size S , used to aggregate nucleotides that will be exploited encode a each pixel and (iii) the transformation function T , which receives, as a parameter, any portion of size S of the genome and transforms it into an integer representing a color. By following this methodology, we can generate a square image in 2 dimensions representing any genome provided as input.

3 Results

In our experiments, the approach has been used, with 4 transformation functions: (i) The *Chargaff index*, (ii) The *Skew index*, (iii) The *Component index*, (iv) The *Diversity index*. The datasets used come from the NCBI (*National Library Medicine*) database. The genome used is a *Mycobacterium tuberculosis* genome⁵ containing 6 194 963 nucleobases.

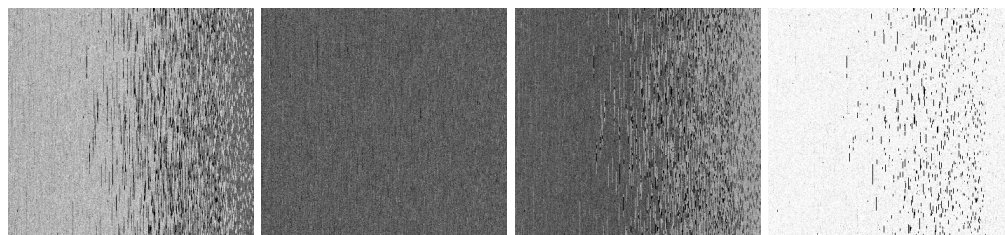


Fig. 1. Images extracted with the four transformation functions on a *Mycobacterium tuberculosis*

4 Conclusion

We have propose an approach that can extract images from genomes. As a perspective for our next works, we want to assess the predictive value of these images and use them for classification purposes.

References

- [1] James D Watson. The human genome project: past, present, and future. *Science*, 248(4951):44–49, 1990.
- [2] Anton Bankevich, Sergey Nurk, Dmitry Antipov, Alexey A. Gurevich, Mikhail Dvorkin, Alexander S. Kulikov, Valery M. Lesin, Sergey I. Nikolenko, Son Pham, Andrey D. Prjibelski, Alexey V. Pyshkin, Alexander V. Sirotkin, Nikolay Vyahhi, Glenn Tesler, Max A. Alekseyev, and Pavel A. Pevzner. Spades: A new genome assembly algorithm and its applications to single-cell sequencing. *Journal of Computational Biology*, 19(5):455–477, 2012. PMID: 22506599.

5. *Mycobacterium tuberculosis* 6505_1: https://www.ncbi.nlm.nih.gov/data-hub/genome/GCA_001318425.1/

BeeDeeM: a general-purpose bioinformatics databank manager system

Patrick Durand

Ifremer, IRSI, SeBiMER Service de Bioinformatique de l'Ifremer, F-29280 Plouzané, France

Corresponding Author: pgdurand@ifremer.fr

Managing sequence databanks is a tedious task since it requires to download data files from many different sources, to handle many different data formats, to prepare various additional indexes to use these banks with bioinformatics tools and to maintain up to date all these banks over time. Data processing also requires to handle in a very efficient way large number of files and disk space, *e.g.* several Tb and thousands of files to deploy banks such as UniprotKB, ENA, Refseq Genomes and NCBI_nt/nr together on a same bioinformatics computing infrastructure.

Very few tools exist to facilitate these tasks. One can cite BioMAJ [1] to handle sequence bank management from the command-line and Galaxy Data Managers [2] to deploy banks for the Galaxy web portals.

To facilitate sequence databanks installation and indexing, we have developed BeeDeeM, a general-purpose bioinformatics databank manager. It provides a suite of command-line and UI tools to manage and enable the easy use of major sequence databanks and biological ontologies. It does not require any additional dependencies (such as MongoDB database for BioMAJ), nor a particular software platform (such as a Galaxy server to use Data Managers).

BeeDeeM implements a modular workflow that automatically performs: (1) the parallel download of the database files from remote sites (via FTP, HTTP or Aspera), (2) the decompression of the files (gzip files), (3) the un-archiving of the files (tar files), (4) the conversion of standard sequence banks (*e.g.* Genbank, Uniprot) to annotated FASTA files, (5) the preparation of databases in BLAST format from native sequence bank formats, (6) the preparation of other indexes such as Diamond, Bowtie, Hisat, *etc.* (7) the indexing of Genbank, Refseq, Embl, Genpept, Swissprot, TrEmbl and Fasta files allowing their efficient querying by way of sequence identifiers, (8) the indexing of sequence features and ontologies data (NCBI Taxonomy, Gene Ontology, Enzyme Commission, Intepro domains and PFAM domains). *BeeDeeM* enables the preparation of taxonomic subsets out of annotated sequence banks and the filtering of sequence banks with user-defined constraints. The software comes with a cluster mode including a dedicated task execution extension: (1) any kind of pre- and post-processing of data can be done using external scripts, (2) such scripts can be executed on the host computer (local mode) or through SGE, PBS or SLURM scheduler (cluster mode). Tasks execution is controlled by configuration files used to specify resources (RAM, CPU, walltime), to setup execution of external softwares (direct execution or through Conda), *etc.*

BeeDeeM being capable of managing annotations available from some data sources (*e.g.* UniprotKB), it comes with a tool capable of annotating BLAST results with very useful information: features table; NCBI Taxonomy; GO, IP, EC terms, *etc.* [3]. Among others, this feature can be used for functional sequence analysis projects.

BeeDeeM is a free and open source Java software. To be as FAIR as possible, the software is also available as a Conda package, a Docker or a Singularity image [4]. This way, *BeeDeeM* can easily be included in other tools, such as Nextflow pipelines (*e.g.* ORSON [5]). It is worth noting that *BeeDeeM* can generate standard '.loc' files enabling the use of banks with a Galaxy server.

References.

1. Filangi, Olivier et al. "BioMAJ: a flexible framework for databanks synchronization and processing." *Bioinformatics* (Oxford, England) vol. 24,16 (2008): 1823-5. doi:10.1093/bioinformatics/btn325
2. <https://galaxyproject.org/admin/tools/data-managers/>
3. <https://pgdurand.gitbook.io/beedeem/utills/cmdline-annotate>
4. <https://github.com/pgdurand/BeeDeeM> ; <https://pgdurand.gitbooks.io/beedeem/>
5. <https://gitlab.ifremer.fr/bioinfo/workflows/orson>

Characterization of particulate matter in the Caribbean area

Lovely Euphrasie-Clotilde¹, Thomas Plocoste^{1,2*}

¹Department of Research in Geoscience, KaruSphère SASU, Abymes 97139, Guadeloupe (F.W.I.)

²Univ Antilles, LaRGE Laboratoire de Recherche en Géosciences et Energies (EA 4539), F-97100 Pointe-à-Pitre, Guadeloupe (F.W.I.)

*Correspondence: thomas.plocoste@karusphere.com

Abstract

The Caribbean area, made up of several archipelagos, is frequently subject to the passages of sands air masses from the African coasts [1]. This natural pollution adds to the aerosol regimes of the Lesser Antilles arc [2]. The latter are mainly composed of marine aerosols, particle linked to volcanic and local anthropogenic activities. Many measures canals are located on the islands of Guadeloupe, Martinique, Cuba, Puerto Rico, and Barbados. More specifically the Aeronet network and ground-based measurements of particulate matter with aerodynamic diameters less than or equal to 2.5 and 10 μm (PM_{2.5} and PM₁₀). Thus, the study of PM_{2.5}/PM₁₀ ratio combined with the optical analysis of the Volume Particle Size Distribution (VPSD) allowed to refine the characterization of the pollution while measuring the induced health impact. Indeed, the PM_{2.5}/PM₁₀ ratio is an indicator relating to the particulate pollution in the atmospheric boundary layer and is also a health barometer. Furthermore, the VPSD profiles highlight the type of particle linked to PM_{2.5}/PM₁₀ ratios identified.

References

- [1] Euphrasie-Clotilde, L., Plocoste, T., Feuillard, T., Velasco-Merino, C., Mateos, D., Toledano, C., Brute, F.-N., Bassette, C., Gobinddass, M., 2020. Assessment of a new detection threshold for PM₁₀ concentrations linked to African dust events in the Caribbean Basin. *Atmospheric Environment*. 117354.
- [2] Plocoste, T., Carmona-Cabezas, R., Jiménez-Hornero, F.J., Gutiérrez de Ravé, E., 2021. Background PM₁₀ atmosphere: In the seek of a multifractal characterization using complex networks. *Journal of Aerosol Science*. 155, 105777.

Equine individual limits for monitoring IGF1 levels in plasma: implementation to the Equine Biological Passport

Agnès BARNABÉ¹, Benoit LOUP¹, Adam CAWLEY², Vivian DELCOURT¹, Patrice GARCIA¹,
Marie-Agnès POPOT¹ and Ludovic BAILLY-CHOURIBERRY¹

¹ GIE LCH, Laboratoire des Courses Hippiques, 15 rue de Paradis, 91370, Verrières-le-Buisson, France

² Australian Racing Forensic Laboratory, Racing NSW, Sydney, NSW 2000, Australia

Corresponding author: a.barnabe@lchfrance.fr

The drug abuse prohibition in races and sports competitions leads to post-event and in-training controls. In horseracing, doping control aims to preserve racing integrity and a relevant selection of breeders based on their performances. Thus, it is essential to monitor the profile of the athletes throughout their careers, especially for the best horses which are considered potential future breeders. To this end and inspired by the human passport implemented in 2008, the French equine doping control laboratory (GIE LCH) and the Racing NSW laboratory (Australia) both integrate an Equine Biological Passport (EBP) for trotters and thoroughbreds respectively, which consists in a longitudinal follow-up of the athletes during a season of competition [1,2]. Based on their prize money performances during the previous racing season, ten French trotters are selected and sampled monthly to monitor a selection of biological and chemical parameters grouped by category as growth hormone modulators.

Growth hormone (GH), also called somatotropin, is naturally secreted in mammals to stimulate the growth and proliferation of tissues. Therefore, recombinant GH (rGH) administration can enhance performance by reducing fat, increasing muscle mass, and expanding physical abilities. One way to control any potential administration is to search for GH's associated marker(s). Several strategies were developed, such as anti-GH antibodies detection, metabolomic fingerprints, and indirect biomarkers detection. The Insulin-like growth factor-I (IGF-1), which could also be potentially misused, has proved its efficiency as an indirect biomarker to detect rGH administration in both human and equine athletes [3]. In fact, the action of GH on the liver upregulates the IGF-1 concentration that tends to remain stable in non-treated animals.

With this method, doping control laboratories can screen for any misuse of rGH in equine athlete samples. However, some covariables, such as gender or age, can impact IGF-1 plasma concentrations. Therefore, reference subpopulations of trotters from the French EBP were built based on a statistical analysis highlighting the impacting covariables. Individual limits are then calculated based on a Bayesian statistical model [4] using the reference subpopulations and the athlete history to monitor the stability of each horse profile. These limits were applied to experimental data after rGH administrations and to data from the Australian EBP to test the relevance of the reference subpopulations. Finally, individual limits were integrated into the new EBP web interface to add specific limits for each horse based on the subpopulation to which it belongs. The results obtained from the application to both EBPs and rGH administration data demonstrate that individual limits constitute a complementary approach to conventional screening strategies.

References

- [1] Arnaud Duluard, Ludovic Bailly-Chouriberry, Fanny Kieken, Marie-Agnès Popot, and Yves Bonnaire. Longitudinal follow-up on racehorses: Veterinary and analytical issues. A one-year study of French trotters. In *Proceedings of the 18th International Conference of Racing Analysts and Veterinarians (ICRAV New Zealand)*, pages 27–34, 2010.
- [2] Adam T. Cawley and John Keledjian. Intelligence-based anti-doping from an equine biological passport. *Drug Testing Analysis*, (9):1441–1447, 2017.
- [3] M.A. Popot, S. Bobin, Y. Bonnaire, P.H. Delahaut, and J. Closset. IGF -I plasma concentrations in non-treated horses and horses administered with methionyl equine somatotropin. *Research in Veterinary Science*, 71(3):167–173, December 2001.
- [4] Martin W. McIntosh, Nicole Urban, and Beth Karlan. Generating longitudinal screening algorithms using novel biomarkers for disease. *Cancer Epidemiol Biomarkers Prev*, 11(2):159–166, February 2002.

Exploring Immune and Tumor Cells in Gliomas Highly Infiltrated by Lymphocytes through Single-Cell RNA-seq and Single-Cell CITE-seq Data Analysis

Jovana Bročić¹, Julie Lerond², Emeline Mundwiller³, Franck Bielle², Justine Guégan¹

¹ Data Analyses Core, Paris Brain Institute, Hôpital de la Pitié-Salpêtrière, 47 Bd de l'Hôpital, 75013 Paris, France

² Genetic and Development of Brain Tumors team, M Sanson/E Huillard, Paris Brain Institute, Hôpital de la Pitié-Salpêtrière, 47 Bd de l'Hôpital, 75013 Paris, France

³ iGenSeq Platform, Paris Brain Institute, Hôpital de la Pitié-Salpêtrière, 47 Bd de l'Hôpital, 75013 Paris, France

Corresponding Author: justine.guegan@icm-institute.org

Gliomas are one of the most lethal types of brain tumors, and immune cells such as lymphocytes have been shown to play a crucial role in their pathogenesis. Since all previous clinical trials and therapeutic approaches did not show any success in glioma patients, there is a need for further research and the development of new therapies [1].

This study is conducted on four samples collected from patients with different histo-pathological glioma subtypes, all with unusually high lymphocytes infiltration. Fluorescence Activated Cell Sorting (FACS) and 10x Chromium experiment were followed by Illumina Next Generation Sequencing (NGS). We used single-cell RNA-seq (scRNA-seq) and single-cell CITE-seq data to investigate the immune and tumor cells (scCITE-seq).

After quality control steps, filtering and normalization, scRNA-seq and scCITE-seq data were integrated into one assay with Weighted Nearest Neighbor (WNN) method and analyzed in R by using Seurat package, version 4.1.0 [2]. Different public data sets - GBM [3] and PBMC [4] - were tested for cluster annotation, but final annotation was set manually according to gene expression levels.

We identified distinct subpopulations of immune and tumor cells based on gene expression profiles and cell surface marker expression, including CD8 T cells, CD4 T cells, Treg CD4 cells, natural killer cells, macrophages and tumor cells.

The detection of rare cell types may provide significant information for diagnosis and treatment. At the moment, we are exploring the interaction between immune and tumor cells, revealing potential regulatory mechanisms that could influence the tumor microenvironment.

Acknowledgements

Jérôme VAN WASSENHOVE, Aurélie GESTIN, Florence DEKNUYDT, Cyto-Ican platform

References

1. Touat M, Li YY, Boynton AN, et al. Mechanisms and therapeutic implications of hypermutation in gliomas, *Nature*, 580(7804):517-523, 2020.
2. Hao Y, Hao S, Andersen-Nissen E. et al. Integrated analysis of multimodal single-cell data, *Cell*, 24; 184(13):3573-3587, 2021.
3. Pombo Antunes, A.R., Scheyltjens, I., Lodi, F. *et al.* Single-cell profiling of myeloid cells in glioblastoma across species and disease stage reveals macrophage competition and specialization. *Nat Neurosci* (24):595–610, 2021.
4. 3k PBMCs from a Healthy Donor (v1, 150x150), Single Cell Gene Expression Dataset by Cell Ranger 1.1.0, 10x Genomics, (2016, May 26).

FliesDB a long story of genomic interface.

Franck Samson*¹ and Carène Rizzon¹

¹Laboratoire de Mathématiques et Modélisation d'Evry – Université d'Évry-Val-d'Essonne, ENSIIE, Université Paris-Saclay, Centre National de la Recherche Scientifique, Institut National de Recherche pour l'Agriculture, l'Alimentation et l'Environnement – France

Résumé

Fliesdb is a genome visualization tool. Our poster will trace the evolution of this tool since its creation and the evolution of the techniques adopted for its development. The technics have changed not only by fashion effect but also due to the obsolescence of language or modifications in the condition of uses. To continue to exist, it is necessary to constantly renew the technology used.

Mots-Clés: web interface, genome, database

*Intervenant

GenoFig: A user-friendly application for the visualization and comparison of genomic regions

Maxime Branger and Sébastien Leclercq*†¹

¹Infectiologie et Santé Publique – Université de Tours, Institut National de Recherche pour l’Agriculture, l’Alimentation et l’Environnement : UMR1282, Institut National de Recherche pour l’Agriculture, l’Alimentation et l’Environnement – France

Résumé

Understanding the molecular evolutionary history of species usually requires visual comparison of genomic regions from related species or strains. Here we present GenoFig, a graphical application for the generation of comparative genomics figures, intended to be as easy to use as possible for biologists and flexible enough to adapt to a variety of needs. GenoFig allows the personalized representation of annotations extracted from GenBank files in a consistent way across sequences, using regular expressions. It also provides several unique options to optimize the display of homologous regions between sequences, as well as other more classical features such as sequence GC percent or GC-skew representations. In summary, GenoFig is a simple, free, and highly configurable tool to explore the evolution of specific genomic regions and to produce publication-ready figures.

Mots-Clés: genomic comparison, visualization, GUI

*Intervenant

†Auteur correspondant: sebastien.leclercq@inrae.fr

PanExplorer: a web application for exploratory analysis and visualization of microbial pan-genomes

Alexis Dereeper^{1,2,*}, Sébastien Cunnac¹, Ralf Koebnik¹, Marilyne Summo^{3,2}, Damien F. Meyer^{4,5*}

(1) PHIM Plant Health Institute, Univ Montpellier, IRD, CIRAD, INRAE, Institut Agro, Montpellier, France

(2) French Institute of Bioinformatics (IFB) - South Green Bioinformatics Platform, Bioversity, CIRAD, INRAE, IRD, 34398 Montpellier, France

(3) CIRAD, UMR AGAP, 34398 Montpellier, France

(4) CIRAD, UMR ASTRE, Centre for Research and surveillance on Vector-borne diseases in the Caribbean, 97170 Petit-Bourg, Guadeloupe, France

(5) ASTRE, CIRAD, INRAE, Univ Montpellier, Montpellier, France

* email: alexis.dereeper@ird.fr, damien.meyer@cirad.fr

In the past decade, the pan-genome concept has been largely employed to investigate the bacterial comparative genomics and evolution analyses. Many programs have been developed for this purpose but a need is still present for the exploration and visualization of data derived from pan-genome analyses. To address this question, we developed a web-based application, PanExplorer, which will perform pan-genome analysis based on the PGAP pipeline and expose resulting information as a comprehensive and easy way, through several modules facilitating the exploration gene clusters and interpretation of data.

The application allows interactive data exploration at different levels :

- Pan-genome visualization as a presence/absence heatmap. This overview allows to easily identify and distinguish core-genes (present in all strains), cloud genes (genes from the accessory genome) and genome-specific genes.
- Physical map of core-genes and strain-specific genes can be displayed as a circular genomic representation (Circos), for each genome taken independently.
- Synteny analysis. The conservation of gene order between genomes can be investigated using graphical representations
- Visual inspection of a specific cluster.

Thanks to the use of NCBI Entrez API, the application guarantees an up-to-date availability of public genomes, to be analyzed on-the-fly, and represents a versatile tool for genome exploration filling a need for bacteriologist community. Among perspectives and further development, new functionalities might be implemented shortly such as on-line pan-genome wide association studies (pan-GWAS) based on Scoary software or COG enrichment statistical studies.

PanExplorer is written in Perl CGI and relies on several Javascript libraries for visualization. It is freely available at <http://panexplorer.southgreen.fr>.

ToolDirectory: Dynamic visualization of softwares managed by Bioinformatics Core Facilities

Alexandre Cormier¹ and Patrick Durand¹

¹ Ifremer, Service de Bioinformatique de l'Ifremer, Brest, France

Corresponding Author: alexandre.cormier@ifremer.fr

Bioinformatics tool management is a crucial task for bioinformatics platforms to have a comprehensive view of all tools available on an infrastructure. It should enable administrators and users to easily navigate through the tool catalog and save time finding appropriate tools. This requires maintaining a list of software in a simple and automatic way after a new installation. In addition, in order to allow end-users to quickly find a tool that fits their needs, it must allow a description and a categorization of each software by the use of various metadata. The latter will improve software classification and allow search and navigation via other criteria, such as the packaging of the tool (compiled source code, Conda package, Singularity image, *etc.*), the operations allowed by the tool (assembly, search for sequence similarity, mapping, *etc.*) or the life science subject (sequence assembly, sequence analysis, epigenetics, phylogenetics).

To meet this need, we have developed ToolDirectory, a Python3 cataloguing and description system for bioinformatics tools available on a computing platform. ToolDirectory allows to describe the many tools and their installed versions using standardized metadata. The information of each tool is stored in a JSON file containing the information specific to the installation (tool name, version, timestamp, packaging, *etc.*) and a comprehensive description (short abstract, EDAM [1] operation, EDAM topic, link to official documentation). To retrieve this metadata, ToolDirectory relies on bio.tools [2], a community-driven curation effort, supported by ELIXIR, that aspires to a comprehensive and consistent registry of information about bioinformatics resources.

ToolDirectory allows you to export in CSV format a subset of the data stored in JSON files to setup their display in a web browser using Katalog JavaScript library. Katalog relies on the faceted viewing system called Keshif (<https://github.com/adilyalcin/Keshif>), on which we have added additional elements to display a catalog of softwares. The different facets are associated to software metadata and make it possible to filter tools very quickly allowing a more efficient and multi-criteria search.

ToolDirectory is open-source package available at <https://github.com/ifremer-bioinformatics/ToolDirectory>.

References

1. Ison, J. et al. EDAM: an ontology of bioinformatics operations, types of data and identifiers, topics and formats. *Bioinformatics* 29, 1325–1332, 2013.
2. Ison, J. et al. Tools and data services registry: a community effort to document bioinformatics resources. *Nucleic Acids Res*, 44, D38–D47, 2016.

A multi-block approach to improve deconvolution of cancer omic data

Elise AMBLARD¹, Vadim BERTRAND¹ and Magali RICHARD¹
CNRS, UMR 5525, VetAgro Sup, Grenoble INP, TIMC, University Grenoble Alpes, 38000 Grenoble, France

Corresponding author: elise.amblard@univ-grenoble-alpes.fr

1 Context

The transcriptome and epigenome are routinely measured in the clinic to diagnose and classify cancer patients. However, these classifications only consider the most abundant tumor subtype in the sample and disregard the others. In contrast, they do not account for intra-tumor heterogeneity, i.e. proportions of the different cell types displayed in a sample. This information is critical as it has an impact on the tumor behavior with respect to its evolution and treatment response.

The current state of the art to de-mix samples is to use one block of data, either transcriptomic or epigenomic. In our project, we would like to combine both blocks. Our work hypothesis is that joint deconvolution should perform better than the current simple-block approach. Indeed, we hope that more information, and from different nature, would improve the de-mixing task.

This project poses the following challenges: *How to do joint deconvolution? Do we improve the de-mixing task with the multi-block approach compared to the simple-block one?*

2 Multi-block strategies

2.1 The deconvolution problem

Let \mathbf{D} be a matrix of size (m, n) with n samples and m molecular features (being methylation probes or transcripts in our case). It can be decomposed into the following product:

$$\mathbf{D} = \mathbf{T} \times \mathbf{A}$$

with \mathbf{T} the matrix of reference profiles displaying the molecular phenotype of k cell types, and \mathbf{A} is the mixing matrix displaying the proportions of the k cell types in the n samples.

There are two main categories of deconvolution tools: supervised or unsupervised. In the supervised approach, the algorithm computes \mathbf{A} (or \mathbf{T}), given \mathbf{D} and \mathbf{T} (or \mathbf{A}). In the unsupervised one, the algorithm computes \mathbf{A} and \mathbf{T} knowing only \mathbf{D} .

We picked 10 state-of-the-art simple-block methods from the literature that we will use as baseline to compare the performance of our multi-block methods.

2.2 Integration and deconvolution of multi-omic data: the theory

When dealing with several -omics from the same samples, the integration step can be done either in an early, intermediary, or late fashion in the analysis pipeline. We plan to test all timings:

Early integration

- ✓ Merge the \mathbf{D} matrices
- ✓ Single initialization

Middle integration

- ✓ Joint dimension reductions
- Alternated initializations
- Optimal transport

Late integration

- ✓ Mean of the $\hat{\mathbf{A}}$'s predicted by each block
- Multi-task learning

2.3 First results and perspectives

So far, we have designed and tested 7 deconvolution tools, declined in 27 settings (designated by [✓] in the list above). We computed an overall rank recapitulating various performance scores to compare methods. The multi-block achieved slightly better results for the supervised case, and positively better for the unsupervised one.

We intend to further explore our ranking process in order to elaborate on the meaning of a score increment. We also plan to design more deconvolution tools (designated by [→] in the list above).

ABEILLE: a novel method for ABerrant Ex-pression Identification empLoying machine Learning from RNA-sequencing data

Justine Labory*^{1,2}, Gwendal Le Bideau , David Pratella , Jean-Elisée Yao , Samira Ait-El-Mkadem Saadi , Sylvie Bannwarth , Loubna El-Hami , Véronique Paquis-Fluckinger , and Silvia Bottini[†]

¹Maison de la Modélisation, de la Simulation et des Interactions [Sophia-Antipolis] – Université Côte d’Azur – 1361 Route des Lucioles 06560 Valbonne, France, France

²Medical Data Laboratory – MDLab - MSI - Université Cote d’Azur – France

Résumé

Current advances in omics technologies are paving the diagnosis of rare diseases proposing as a complementary assay to identify the responsible gene. The use of transcriptomic data to identify aberrant gene expression (AGE) have demonstrated to yield potential pathogenic events. However popular approaches for AGE identification are limited by the use of statistical tests that imply the choice of arbitrary cut-off for significance assessment and the availability of several replicates not always possible in clinical contexts. Machine learning methods via neural networks including autoencoders (AEs) or variational autoencoders (VAEs) have shown promising performances in medical fields.

Here, we describe ABEILLE, (ABerrant Expression Identification empLoying machine LEarning from sequencing data), a novel method for the identification of AGE from RNA-seq data without the need of replicates or a control group, using a flexible model obtained after testing several parameters. ABEILLE combines the use of a VAE, able to model any data without specific assumptions on their distribution, and a decision tree to classify genes as AGE or non-AGE. An anomaly score is associated to each AGE in order to stratify them by severity of aberration.

We compare ABEILLE performances to the state-of-the-art alternatives by using semi-synthetic data and a real dataset demonstrating the importance of the flexibility of the VAE configuration to identify potential pathogenic candidates.

Mots-Clés: Rare disease, Transcriptomics, Variational autoencoder

*Intervenant

[†]Auteur correspondant: Silvia.BOTTINI@univ-cotedazur.fr

adverSCarial: a tool for evaluating adversarial attacks on single-cell transcriptomics classifiers

Ghislain Fievet and Sébastien Hergalant

UMRS INSERM 1256 NGERE, 9 Av. de la Forêt de Haye, 54500, Vandœuvre-lès-Nancy, France

Corresponding Author: ghislain.fievet@univ-lorraine.fr

In single-cell transcriptomics, machine learning techniques have been applied for automatic cell annotation [1], for the identification of cancer cell subpopulations [2], and for modelling the transcriptional dynamics that govern cellular fate and development [3-4]. These methods hold potential value for routine practice in clinical settings but must address critical challenges in the use – and misuse – of AI algorithms for reliability and interpretability [5]. The field of explainable AI addresses these concerns, among which the robustness to adversarial attacks, *i.e.* techniques designed to fool a machine learning model with deceptive and inaccurate data.

Here we present *adverSCarial*, an original R package that generates adversarial attacks on single-cell transcriptomics classifiers (<https://github.com/GhislainFievet/adverSCarial>). *adverSCarial* currently proposes four customizable functions to produce adversarial attacks. In this work, we define two types of methods: the minimal (min) and the maximal (max) change attacks. The min change attack finds the smallest possible perturbation in the input data leading to a change of classification. The max change attack finds the largest data modification which does not alter the initial classification.

On a reference peripheral blood mononuclear cells (PBMC) dataset of 2,700 cells and 22,042 genes [6], we further tested *adverSCarial* and compared the susceptibility of two published cell type classifiers to these attacks, the classification tree based CHETAH [7] and the marker based scType [8]. Both classifiers showed weaknesses to min and max adversarial attacks, especially to the max change method. Indeed, we found that it is possible to modify a significant proportion of genes without affecting the classification confidence. For a representative example, on the CD14 monocytes cluster of the studied PBMC dataset, replacing all 22,042 gene expression values by their last percentile did not modify CHETAH identification as CD14 cells. These results were generalized to all cell types and highlight the concern that machine learning algorithms may fail to detect even significant anomalies in the input data.

In conclusion, this work demonstrates the usefulness of such techniques in testing the robustness of classifier families and the extent of the modifications required to deviate their intended use. We believe that our approach and tool can guide the development and validation of more reliable models that could be used in clinical setups, and aim to extensively evaluate these tools on a wide variety of single-cell datasets.

References

- [1] Xinlei Zhao, Shuang Wu, Nan Fang, Xiao Sun, Jue Fan. Evaluation of single-cell classifiers for single-cell RNA sequencing data sets. *Briefings in Bioinformatics*, pages 1581-1595, 2020.
- [2] Dohmen, J., Baranovskii, A., Ronen, J. et al. Identifying tumor cells at the single-cell level using machine learning. *Genome Biol*, 2022.
- [3] Kushagra Pandey, Hamim Zafar. Inference of cell state transitions and cell fate plasticity from single-cell with MARGARET. *Nucleic Acids Research*, 2022.
- [4] Saelens, W., Cannoodt, R., Todorov, H. et al. A comparison of single-cell trajectory inference methods. *Nat Biotechnol*, 2019.
- [5] de Hond, A.A.H., Leeuwenberg, A.M., Hooft, L. et al. Guidelines and quality criteria for artificial intelligence-based prediction models in healthcare: a scoping review. *npj Digit. Med.*, 2022.
- [6] Zheng, G., Terry, J., Belgrader, P. et al. Massively parallel digital transcriptional profiling of single cells. *Nat Commun*, 2017.
- [7] Jurriaan K de Kanter, Philip Lijnzaad, Tito Candelli, Thanasis Margaritis, Frank C P Holstege, CHETAH: a selective, hierarchical cell type identification method for single-cell RNA sequencing. *Nucleic Acids Research*, 2019.
- [8] Ianevski, A., Giri, A.K. & Aittokallio, T. Fully-automated and ultra-fast cell-type identification using specific marker combinations from single-cell transcriptomic data. *Nat Commun*, 2022.

An automated PVC/SR QRS discriminator on 12-lead ECG

Amèle AHRAOUI^{1,2} and Nicolas CEDILNIK²

¹ Université Côte d’Azur, 2004 Rte des Lucioles, 06902, Valbonne, France

² InHeart, IHU Liryc - Hôpital Xavier Arnoz Avenue du Haut Lévêque, 33600 Pessac France

Corresponding author: amele.ahraoui@inheartmedical.com

1 Context

Idiopathic premature ventricular contractions (PVC) are an arrhythmia treatable by catheter ablation [1], which destroys their source. However, PVCs are not always observed during the ablation intervention, and locating their source often rely on prior electrocardiography (ECG) analysis by a cardiologist. In this work, we explore automation of this human analysis with machine learning; as a first step, we set out to build a classifier that can discriminate sinus rhythm (SR) from PVC QRS complexes.

2 Data and methods

First, we processed the data [2] by correcting the wandering baseline using a FIR filter applied to our ECG signals [3]. Then, using the Biosppy library, we detected R-peaks following the approach by Hamilton. The QRS window was defined by using empirically choosing 10% of the R-R distance. We extracted 3 features from these QRS: the minimum and maximum voltage per derivation and the QRS area under the curve, for a total of 36 features per QRS complex. We trained multiple classifiers and evaluate their performances with a 5-fold cross-validation on a public database [4] where we annotated 333 PVCs and 333 SRs.

3 Results and conclusion

Our QRS features are relevant for distinguishing between PVC and SR complexes, as most classifiers performed with a F1 Score over 0.97 (Fig.1). A qualitative review of the predictions made on an unlabelled database [2] confirmed that performance seemed acceptable. Some errors are visible (see Fig.2), most likely due to a data/feature normalization and QRS windowing issues. Our future work will focus on solving these issues.

Classifiers	Accuracy	F1 Score
Nearest Neighbors	0.985	0.985
Linear SVM	0.986	0.986
Gaussian Process	0.991	0.991
Decision Tree	0.974	0.974
Random Forest	0.979	0.979
Neural Net	0.992	0.992
AdaBoost	0.986	0.986
Naive Bayes	0.982	0.982
QDA	0.979	0.979

Fig. 1. Performance of several classifiers. Differences are likely not significant.

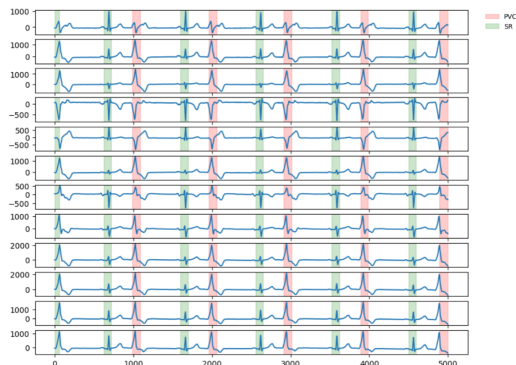


Fig. 2. Prediction (neural net) on an ECG sample picked from [2].

References

- [1] Rakesh Latchamsetty et al. Multicenter outcomes for catheter ablation of idiopathic premature ventricular complexes. *JACC: Clinical Electrophysiology*, 1(3):116–123, 2015.
- [2] Perez Alday et al. Classification of 12-lead ECGs: the PhysioNet/Computing in Cardiology Challenge 2020. *Physiological Measurement*, 41(12):124003, December 2020.
- [3] Victor de Pinto. Filters for the reduction of baseline wander and muscle artifact in the ECG. *Journal of Electrocardiology*, 25:40–48, 1992.
- [4] Jianwei Zheng et al. A 12-lead ecg database to identify origins of idiopathic ventricular arrhythmia containing 334 patients. *Scientific Data*, 7(1):98, 2020.

ECGtizer: a tool for paper ECG digitization

Alex Lence^{1,*}, Ahmad Fall², Jean-Daniel Zucker^{1,2} and Edi Prifti^{1,2}

¹IRD, Sorbonne Université, Unité de Modélisation Mathématique et Informatique des Systèmes Complexes, UMMISCO, F-93143, Bondy, France

²Sorbonne Université, INSERM, Nutrition et Obésités; systemic approaches, NutriOmique, AP-HP, Hôpital Pitié-Salpêtrière

*alex.lence@ird.fr

Tools for digitizing paper ECGs are currently limited and non-easily accessible. Indeed, for most of them they are semi-automated and do not support different formats of paper ECGs. Here we present ECGtizer, which is a fully automated tool for ECG digitization supporting up to 5 different ECG formats (12 lead Mortara, DuoEK, Pulsbit, Kardia and AppleWatch). Moreover, ECGtizer provides useful functionalities, including anonymization of patient information, signal completion, multi-format printouts, etc. However, the digitisation of ECGs is not enough, as we have realized that paper ECGs only represent part of the information of the recording. In other words, and to give an example, if we take a 10-second recording of 12 leads, each lead will be represented on the paper for only 2.5 seconds. For this reason, we have developed a completion deep learning model to complete each lead on 10sec.

For digitization, we have combined the approaches described in [1] with other approaches to automatically extract the signal from different ECG formats. Our tool then allows to complete the missing parts of the leads. To do this, we trained a U-net to reconstruct the missing parts of each lead. More precisely, we have given ECGs of 12 leads as input, to which we have applied masks, and the U-net has the task of reconstructing the masked parts. To our knowledge, there are only two other approaches to ECG reconstruction copy paste and GAN [2,3].

We obtain a Pearson correlation after digitization of 0.95, which is comparable to [1,2], and an MSE of 0.001mV, which is better than [2] which obtained 0.016mV. Finally, for the completion, we also obtain results better than the state of the art of [3] with a MSE of 0.005mV with our model against 0.03 for copy paste and 0.016 for GAN.

We offer a fully automated, open-source tool with state-of-the-art performance and the ability to complete the missing information for each lead.

References

- [1] M. Baydoun, L. Safatly, O. K. Abou Hassan, H. Ghaziri, A. El Hajj, and H. Isma'eel, "High Precision Digitization of Paper-Based ECG Records: A Step Toward Machine Learning," *IEEE J. Transl. Eng. Health Med.*, vol. 7, pp. 1–8, 2019, doi: 10.1109/JTEHM.2019.2949784.
- [2] F. Badilini, T. Erdem, W. Zareba, and A. J. Moss, "ECGScan: a method for conversion of paper electrocardiographic printouts to digital electrocardiographic files," *J. Electrocardiol.*, vol. 38, no. 4, pp. 310–318, Oct. 2005, doi: 10.1016/j.jelectrocard.2005.04.003.
- [3] H.-C. Seo, G.-W. Yoon, S. Joo, and G.-B. Nam, "Multiple electrocardiogram generator with single-lead electrocardiogram," *Comput. Methods Programs Biomed.*, vol. 221, p. 106858, Jun. 2022, doi: 10.1016/j.cmpb.2022.106858.

Inference of biological interactions on large heterogeneous graph networks with machine learning

Antoine TOFFANO¹, Jacques PÉCREAUX¹ and Christophe HÉLIGON¹
CNRS, Univ Rennes, IGDR - UMR 6290, 2 avenue du Professeur Léon Bernard, F-35000, Rennes, France

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

Finding candidate genes remains key to deciphering molecular mechanisms in biology and medicine. It is often based on the predicted relationships between genes and phenotypes. However, deriving those connections from the ever-growing amount of biological data and knowledge is becoming increasingly challenging for individuals. This complex and heterogeneous data can be modeled in computer-actionable graphs using frameworks like Semantic Web technologies.

The present work aims to take advantage of multiple data types to discover new relations between biological entities, like genes, phenotypes or diseases. We foresee that combining heterogeneous data graphs and machine learning will enable us to relate those biological entities in a manner resilient to both sparse and noisy data.

In this context, the nematode *C. elegans*, a classic model organism, provides a rich data source thanks to in-depth genetics studies and a vast array of available experimental data. Furthermore, the WormBase database already gather most of this information on *C. elegans*, namely genes, phenotypes, interactions between genes, diseases, and gene expression patterns. We represent this data as a large heterogeneous knowledge graph and analyze it using graph machine learning methods. Since different algorithms can capture different aspects of the graph, we set to benchmark both foundational and state-of-the-art machine learning algorithms.

Ours is a two-step methodology. We first train a graph embedding model that can generate a numerical representation of the nodes and relations. Secondly, we train a classifier that uses those embeddings to determine whether a relation exists between two chosen nodes. To compare the embedding algorithms, we randomly delete 20% of the edges and train the algorithm on the leftover 80%. We then use half of the removed edges as a test set to assess embedding quality. We reach a Hit@1 on a relation prediction task of about 0.95 with ConvKB, a graph convolutional neural network algorithm, although other algorithms can match this performance. Equipped with the most performing algorithm, we aim to predict novel gene-phenotype associations with a binary classifier. We reach an accuracy of 0.82 using a Random Forest in predicting the existence of a link.

In a broader take, our work will shed light on the underlying biological mechanisms linking genes and phenotypes, and will accelerate the cell biology research in *C. elegans*. Our approach being data agnostic can apply to a wide range of both data type and organisms.

Acknowledgements

We thank Prof. Olivier Dameron (Dyliss team, IRISA, Univ Rennes) for sharing his expertise on heterogeneous data integration and mining.

Issues in UK Biobank for GWAS: case-control definition and imbalance

Margot DEROUIN¹, Sidonie FOULON^{1,3}, Marie-Sophie OGLOBLINSKY², Hervé PERDY³ and Anne-Louise LEUTENEGGER¹

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

² Univ Brest, Inserm, EFS, UMR 1078, GGB, F-29200 Brest, France

³ CESP, Inserm, UMR1018, Université Paris Saclay, 16 Avenue Paul-Vaillant-Couturier, 94807, Villejuif, France

Corresponding Author: margot.derouin@inserm.fr

With more than 500,000 individuals, the UK Biobank [1] is one of the largest biobanks in the world and a valuable resource for public health research. The UK Biobank recruited people aged between 40-69 years and living in the United Kingdom to take part in this project in 2006-2010. Genetic information, health and lifestyle data for each participant are available and regularly updated, making a wide range of studies possible.

Initially, the large size of the cohort was based on statistical power calculations for case-control studies making it a major contributor to the advancement of modern medicine and treatment to improve human health. Nevertheless, the richness of this database is a double-edged sword. Dealing with this amount of information turns out to be challenging, especially when it comes to focusing on a specific disease. One issue is how to define whether an individual is a case or a control as different data sources (e.g. self-reported questionnaires, hospital records ...) might provide contradictory information. Another issue is the resulting number of cases vs. controls, which can be highly unbalanced.

Here we choose to illustrate these issues on the type 2 diabetes (~90 percent of all diabetic patients [2]). We show how they can impact the results of the classic additive GWAS and GWAS focusing on rare recessive variants through homozygous-by-descent segments HBD-GWAS using R packages Gaston [3] and Fantasio [4].

Acknowledgements

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

References

- [1] Bycroft C, Freeman C, Petkova D, Band G, Elliott LT, Sharp K, et al. The UK Biobank resource with deep phenotyping and genomic data. *Nature*. 2018 Oct;562(7726):203–9.
- [2] Pei J, Wang B, Wang D. Current Studies on Molecular Mechanisms of Insulin Resistance. *J Diabetes Res*. 2022 Dec 23;2022:1863429. doi: 10.1155/2022/1863429. PMID: 36589630; PMCID: PMC9803571.
- [3] Perdry H, Dandine-Roulland C. *gaston*: Genetic Data Handling (QC, GRM, LD, PCA) & Linear Mixed Models. 2022. R package version 1.5.9, <https://CRAN.R-project.org/package=gaston>
- [4] Derouin M, Foulon S, Leutenegger AL, Perdry H, <https://github.com/genostats/Fantasio2>

Key-words : UK Biobank; GWAS; consanguinity; Homozygosity-by-descent ; type 2 diabetes

Prediction and classification of methylation class of brain CNS tumour

Fabien JOSSAUD^{1,2}, Florent CHUFFART^{1,2}, Anne MCLEER^{1,3}, Jean BOUTONNAT³ and Julien THEVENON^{1,2,3}

¹ IAB, Site Santé, Allée des Alpes, 38700, La Tronche, France

² MIAI, Université Grenoble-Alpes, 38400, Saint Martin d'Hères, France

³ CHU Grenoble-Alpes, Site Santé, Allée des Alpes, 38700, La Tronche, France

Corresponding author: fabien.jossaud@univ-grenoble-alpes.fr, florent.chuffart@univ-grenoble-alpes.fr

Clinical management of central nervous system tumours routinely include methylation profiling. Methylation assay allow the histo-pathological refinement and sub-classification of specific tumours to help the diagnostic, prognosis and therapeutic management of patients. Grenoble-Alpes University Hospital have developed since 2021 a local procedure for exploring tumoral methyloma. Until now they were sub-contracting the bioinformatics and biostatistical analysis to a foreign group.

This work highlights the synergic collaboration of the University-Hospital with the EpiMed research group, with a longstanding experience in methylome data analysis. We have created a dedicated classifier based on public data cumulating 1093 brain tumours (GSE109379), and predicted age comparing several epigenetics clocks, and in-house tumour classification model among the 91 published methylation classes. We compared our results on public data and on local samples. The accuracy of our current classifier on cross-validation training public data was estimated at 98% for gender and 78% for the 91 methylation classes.

The figure below highlights the t-SNE projection of the samples (triangles) on the classification cohort (dots), showing the 91% of accurate classification of tumors according to the 13 super-classes from the literature.

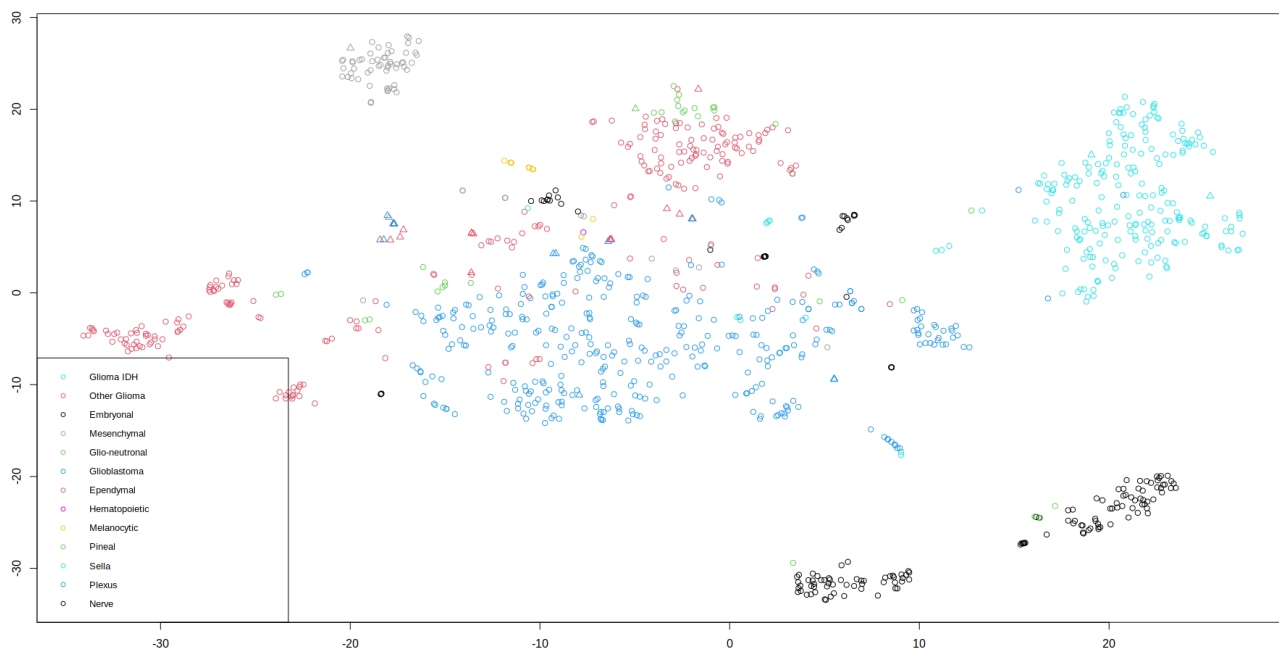


Fig. 1. T-sne methylation super-classes

During the process of generating the classifier, we experienced the need for a finely tuned selection of biological constraints to include in the training model, notably sample type and conservation. In conclusion, our preliminary results suggest that a routine usage of our classifier would help the diagnosis of patients, and provide valuable datasets for basic research.

Random Walk with Restart on multilayer networks: from node prioritization to supervised link prediction and beyond

Anthony BAPTISTA^{1,2,3}, Galadriel BRIÈRE⁴ and Anaïs BAUDOT¹

¹ Aix Marseille Univ, INSERM, MMG, 13385, Marseille, France

² The Alan Turing Institute, The British Library, London, NW1 2DB, United Kingdom

³ School of Mathematical Sciences, Queen Mary University of London, London, E1 4NS, United Kingdom

⁴ Aix Marseille Univ, CNRS, I2M, Marseille, France

Corresponding author: marie-galadriel.briere@univ-amu.fr

Background

In the era of big data, biological networks have proven invaluable for representing biological knowledge gathered from a wide number of sources, including public databases and omics data. Universal multilayer networks, which gather interactions between different types of nodes and edges and combine multiplex, heterogeneous and bipartite networks, provide a natural way to integrate such diverse and multi-scale data sources into a common framework. Recently, we developed MultiXrank, a Random Walk with Restart algorithm able to explore such universal multilayer networks [1]. Starting from one or more seed nodes, the random walker navigates the different network layers and generates scores that reflect node's relevance with respect to the seed(s). These scores can then be used in a wide variety of downstream analyses. We aim here to highlight these versatile usages in various bioinformatics tasks.

Results

Node prioritization First, we show that MultiXrank scores can be directly used for node prioritization. Specifically, using a multilayer network containing interactions between genes, drugs, and diseases, we prioritized genes and drugs associated with leukemia and epilepsy. For drug prioritization in epilepsy, we also show that a simple ranking of nodes based on their MultiXrank scores shows comparable predictive performance as the supervised approach used by Hetionet [2].

Supervised classification In a second study, we show that MultiXrank scores can also be used in a supervised manner for link prediction. We trained a classifier to predict gene-disease associations from MultiXrank scores. The training is done using a multilayer network containing interactions between genes and diseases gathered from an outdated version of DisGeNET (v2.0, 2014) [3]. To test the performance of the classifier, we compared our predictions with the gene-disease associations present in an updated version of the DisGeNET database (v7.0, 2020).

Diffusion profiles comparison Finally, we show that MultiXrank scores can be used to compute diffusion profiles from various seeds. The diffusion profiles can then be considered as signatures and compared. We built a multilayer network containing gene and disease interactions layers, as well as cell-type specific genomic network layers built from PC Hi-C and TAD data from hematopoietic cells. We then selected immune diseases nodes as seeds, and produced a diffusion profile signature for each disease in each cell-type. We observed that those signatures capture the hematopoietic cell lineages. Further comparison of the disease signatures uncover comorbidity relationships between immune diseases.

References

- [1] Anthony Baptista, Aitor Gonzalez, and Anaïs Baudot. Universal multilayer network exploration by random walk with restart. *Communications Physics*, 5(1):1–9, July 2022. Number: 1 Publisher: Nature Publishing Group.
- [2] Daniel Scott Himmelstein, Antoine Lizee, Christine Hessler, Leo Brueggeman, Sabrina L Chen, Dexter Hadley, Ari Green, Pouya Khankhanian, and Sergio E Baranzini. Systematic integration of biomedical knowledge prioritizes drugs for repurposing. *eLife*, 6:e26726, September 2017.
- [3] Janet Piñero, Núria Queralt-Rosinach, Alex Bravo, Jordi Deu-Pons, Anna Bauer-Mehren, Martin Baron, Ferran Sanz, and Laura I. Furlong. Disgenet: a discovery platform for the dynamical exploration of human diseases and their genes. *Database (Oxford)*, 2015, January 2015.

Scoring and ranking strategies to benchmark cell type deconvolution pipelines

Vadim BERTRAND¹, Elise AMBLARD¹ and Magali RICHARD¹
CNRS, UMR 5525, VetAgro Sup, Grenoble INP, TIMC, University Grenoble Alpes, 38000 Grenoble, France

Corresponding author: vadim.bertrand@etu.univ-grenoble-alpes.fr

1 Context

With the emergence of new standards of care for cancer patients, omics data are now routinely collected. These data are necessary for diagnostic purposes, as they allow classification of patients. However, current classifications could be improved, especially by taking into account the cell type composition of the tumor (tumor heterogeneity). Because of the complexity of tumor heterogeneity, there is still no consensus framework for estimating tumor composition, although many deconvolution algorithms have been developed to address this problem over the past decade. In order to evaluate and compare these new deconvolution methods, a robust and comprehensive scoring and ranking strategy is needed. To construct such benchmarks, we built a ranking pipeline based on several evaluation criteria. We used different methods to aggregate the scores and finally established an overall ranking. This pipeline also integrates approaches to assess the significance of differences between the aggregated scores and the stability of our ranking.

2 Benchmark of deconvolution pipelines

2.1 Cell type deconvolution

Estimating cell type proportions from bulk samples can be posed as the following matrix factorization problem:
$$D_{G \times S} = T_{G \times K} \times A_{K \times S}$$

with D being the bulk matrix of S samples and G genes, T a reference profiles matrix of the G genes for K cell types and A the matrix of the K cell type proportions for the S samples.

2.2 Our benchmark setting

In order to benchmark on a large diversity of samples, we use simulated data emulating $N = 3$ different cancer types. For every dataset, we simulate $M = 10$ proportions matrices $A_{n,m}$ by sampling from a Dirichlet distribution - representing the biological noise - and by multiplying them by an in-vitro reference matrix T_n , after adding a technical noise $\epsilon_{n,m}$ we finally obtain $N * M = 30$ simulated bulk matrices $D_{n,m} = T_n \times A_{n,m} + \epsilon_{n,m}$.

2.3 Scoring and ranking strategies

We selected 4 evaluation categories: computational cost, raw performance, stability and consensus, in order to balance deconvolution accuracy with practicability and generalisation. We used various scores within every category, ranging at different intervals. Therefore, constructing a final score requires a normalization step and several aggregations: across simulations of a given dataset, across datasets, across scores within categories and finally across categories. At every step, we tested different aggregation operators, such as median or means (arithmetic, harmonic or geometric).

From final scores, we derived a global ranking yielding the "winning" pipeline. We investigated the validation of several criteria accessing the "fairness" of our ranking methods: majority, Condorcet, consistency, participation and independence of irrelevant alternatives.

In addition to the "fairness" of ranking approaches, it is also of interest to quantify final score differences between pipelines in order to detect significant performance improvement. To this purpose, we computed final p - values by either (i) aggregating intermediate p - values calculated from parametric tests using Stouffer's or Edgington's methods; or (ii) employing a single non-parametric permutation test procedure.

Development of a knowledge graph framework to ease and empower translational approaches in plant research: a use-case study on grain legumes

Baptiste IMBERT¹, Jonathan KREPLAK¹, Raphaël-Gauthier FLORES^{2,3}, Grégoire AUBERT¹, Judith BURSTIN¹

¹ Agroécologie, INRAE, Institut Agro, Univ. Bourgogne, Univ. Bourgogne Franche-Comté, F-21000 Dijon, France

² Université Paris-Saclay, INRAE, URGI, 78026, Versailles, France

³ Université Paris-Saclay, INRAE, BioinfOmics, Plant bioinformatics facility, 78026, Versailles, France

Corresponding Author: baptiste.imbert@inrae.fr
nadim.tayeh@inrae.fr

While the continuing decline in genotyping and sequencing costs has largely benefited plant research, some key species for meeting the challenges of agriculture remain largely understudied. As a result, heterogeneous datasets for different traits are available for a significant number of these species. As gene structures and functions are to some extent conserved through evolution, comparative genomics can be used to transfer available knowledge from one species to another. However, such a translational research approach is complex due to the multiplicity of data sources and the non-harmonized description of the data.

Here, we provide two pipelines, referred to as structural and functional pipelines, to create a framework for a NoSQL graph-database (Neo4j) to integrate and query heterogeneous data from multiple species. We call this framework Orthology-driven knowledge base framework for translational research (Ortho_KB). The structural pipeline builds bridges across species based on orthology. The functional pipeline integrates biological information, including quantitative trait loci (QTL), RNA-seq datasets, and uses the backbone from the structural pipeline to connect orthologs in the database. Queries can be written using the Neo4j Cypher language and can, for instance, lead to identify genes controlling main traits across species. To explore the possibilities offered by such a framework, we populated Ortho_KB to obtain OrthoLegKB, an instance dedicated to legumes. The proposed model was evaluated by studying the conservation of a flowering-promoting gene. Through a series of queries, we have demonstrated that our knowledge graph base provides an intuitive and powerful platform to support research and development programmes.

Poster 3

A 4-Year retrospective study of the presence of thermophilic free-living amoebae in recreative baths in Guadeloupe.

Isaure Quétel*¹, Youri Vingataramin , Didier Roux , Marie-Anne Pons , Antoine Talarmin , and Isabel Marcelino

¹Institut Pasteur de la Guadeloupe, TRed-Path Unit, Les Abymes, Guadeloupe, France – Guadeloupe

Résumé

Free-living amoebae (FLA) are ubiquitous protists found in soil and water across the globe. Some FLA such as *Naegleria fowleri* (*NF*), *Acanthamoeba*, *Sappinia* and *Balamuthia* can cause rare but fatal encephalitis 1. In 2008, *NF* was responsible for the death of a 9-year-old boy who swam in a recreational bath in Guadeloupe 2. In 2013, our group showed that *NF* could be found in most of these baths 3, the soil being the origin of this contamination 4.

In this work, we aimed to analyze the diversity and abundance of thermophilic FLA (and in particular *Naegleria sp*) in recreational waters in Guadeloupe over 4 years, using metabarcoding and the most probable number (MPN) method, and herewith detect their possible seasonality.

From 2018 to 2022, a total of 74 water samples were collected from 7 baths commonly used by Guadeloupean and tourists, and with different characteristics (natural, tiled, regularly cleaned or not, and with temperatures ranging from 27 to 40°C). DNA was extracted from FLA cultivated at 37-40°C to detect thermophilic FLA. Metabarcoding studies were conducted through FLA 18S rRNA amplicons sequencing 5; amplicon sequence variants (ASV) were extracted from each sample and taxonomy assigned against PR2 database using dada2 and phyloseq tools. We also searched for *Naegleria* and *NF* using conventional PCR targeting ITS and NF-ITS 3 and we quantified these FLA using the MPN 6.

Our results showed that, over the 4 years, high differences in FLA diversity and abundance were observed between the 7 baths, but no seasonal distribution was concluded. *Naegleria*, *Vermamoeba* and *Stenamoeba* were the most represented genera, while the genera *Acanthamoeba* and *Vahlkampfia* were mainly found in 2 baths. Furthermore, the MPN values for *Naegleria sp* (NT/L) increased overall between 2018 and 2022 in almost all baths but the MPN values for *NF* (NF /L) seem to decrease.

Globally, our results showed that although we cannot establish a peak in FLA detection, the presence of *Naegleria* and *Acanthamoeba* in recreational baths can pose a potential threat on human health in Guadeloupe. As such, it is important to continue the regular control of these baths.

*Intervenant

Mots-Clés: thermophilic free, living amoebae, 18S metabarcoding, recreational water

A novel and dual digestive symbiosis scales up the nutrition and immune system of the holobiont *Rimicaris exoculata*

Johanne AUBE¹, Marie-Anne CAMBON-BONAVITA¹, Lourdes VELO-SUAREZ^{1,2,3}, Valérie CUEFF-GAUCHARD¹, Françoise LESONGEUR¹, MARION GUEGANTON¹, LUCILE DURAND¹ and Julie REVEILLAUD^{1,4}

¹ Univ Brest, CNRS, Ifremer, UMR6197 Biologie et Ecologie des Ecosystèmes marins Profonds, 29280, Plouzané, France

² Univ Brest, INSERM, EFS, UMR 1078, GGB, Brest, France

³ Centre Brestois d'Analyse du Microbiote (CBAM), Brest University Hospital, 29200, Brest, France

⁴ MIVEGEC, Univ. Montpellier, INRAe, CNRS, IRD, 34398, Montpellier, France

Corresponding Author: johanne.aube@ifremer.fr

In deep-sea hydrothermal vent areas, deprived of light, most animals rely on chemosynthetic symbionts for their nutrition. These symbionts may be located on their cuticle, inside modified organs, or in specialized cells. Nonetheless, many of these animals have an open and functional digestive tract. The vent shrimp *Rimicaris exoculata* is fueled mainly by its gill chamber symbionts, but also has a complete digestive system with symbionts. These are found in the shrimp foregut and midgut, but their roles remain unknown. We used genome-resolved metagenomics on separate foregut and midgut samples, taken from specimens living at three contrasted sites along the Mid-Atlantic Ridge (TAG, Rainbow, and Snake Pit) to reveal their genetic potential.

We reconstructed and studied 20 Metagenome-Assembled Genomes (MAGs), including novel lineages of *Hepatoplasmataceae* and *Deferribacteres*, abundant in the shrimp foregut and midgut, respectively. Although the former showed streamlined reduced genomes capable of using mostly broken-down complex molecules, *Deferribacteres* showed the ability to degrade complex polymers, synthesize vitamins, and encode numerous flagellar and chemotaxis genes for host-symbiont sensing. Both symbionts harbor a diverse set of immune system genes favoring holobiont defense. In addition, *Deferribacteres* were observed to particularly colonize the bacteria-free ectoperitrophic space, in direct contact with the host, elongating but not dividing despite possessing the complete genetic machinery necessary for this.

Overall, these data suggest that these digestive symbionts have key communication and defense roles, which contribute to the overall fitness of the *Rimicaris* holobiont.

Assessing the hidden biodiversity of coral reefs in Guadeloupe using Autonomous Reef Monitoring Structures (ARMS)

Pierre-Louis RAULT¹, Quentin DERIAN¹, Fabio GERBERON¹, Mathias GERENIUS¹, Riwelen LE FOLL¹, Sébastien CORDONNIER¹, Amélia CHATAGNON¹, Josie LAMBOURDIÈRE¹, Françoise DENIS¹, Charlotte DROMARD¹ & Etienne BEZAULT¹

¹ UMR BOREA, Laboratoire Biologie des ORganismes et Ecosystèmes Aquatiques, CNRS, MNHN, IRD, Sorbonne Université, Université de Caen & Université des Antilles, 97.159 Pointe à Pitre, Guadeloupe, F.W.I

Corresponding author: etienne.bezault@univ-antilles.fr

Abstract:

Coral reefs are renowned for their high biodiversity as well as important decline worldwide, but much of this diversity remains hidden or underestimated by traditional monitoring methods. Autonomous Reef Monitoring Structures (ARMS) offer a standardized protocol for assessing the "cryptic" fauna that inhabits the coral reef ecosystems. In this study, we deployed, for the first time in Guadeloupe, ARMS modules at two sites within the St François lagoon, one close to the natural reef and the other close to an eco-mooring site. Our study aimed at comparing the ability of ARMS to evaluate the biodiversity between natural and anthropogenic sites and preparing for a larger-scale study of reef cryptodiversity in the Lesser Antilles region.

After six months of deployment, we retrieved the ARMS modules in order to estimate and identify the colonizing vagile and sessile faunas from both sites, using two quantitative and computational-based approaches. The colonization and diversity of the sessile fauna was estimated using first photo-analysis with randomized sampling methods and second NGS metabarcoding techniques with multimarkers approach.

Based on photo-analysis, preliminary results suggested differences in the community composition between the two study sites (*e.g.*, higher settlement covering rate on the eco-mooring ARMS than on the reef ARMS) and also within the ARMS, depending on micro-environmental factors (*e.g.*, gradient of Coralline algae cover rate, with higher cover at the top and lower cover at the bottom, in contrast to the filtering organisms such as polychaetas). Considering Metabarcoding, literature and genetic database searches were conducted in order to evaluate the optimal protocol and/or gene markers to be used to characterize the more accurately the colonizing biodiversity including metazoans, eukaryotes and prokaryotes.

This integrative approach based on ARMS seems to be of prime interest in both detecting the cryptic biodiversity of coral reefs in Guadeloupe, and further monitor the dynamics of coral reef diversity colonization across the Caribbean islands and/or gradient of anthropization.

Characterization and quantification of antibiotic resistance gene variants in gut microbiota.

Ouléye Sidibé^{*1}, Anne-Carmen Sanchez², Guillaume Kon-Kam-Kim², Fanny Calenge³, Benoît Doublet¹, Sylvie Baucheron¹, Sébastien Leclercq¹, and Anne-Laure Abraham²

¹Infectiologie et Santé Publique (ISP) – Institut National de Recherche pour l’Agriculture, l’Alimentation et l’Environnement – France

²Mathématiques et Informatique Appliquées du Génome à l’Environnement [Jouy-En-Josas] – Institut National de Recherche pour l’Agriculture, l’Alimentation et l’Environnement – France

³Génétique Animale et Biologie Intégrative – AgroParisTech, Université Paris-Saclay, Institut National de Recherche pour l’Agriculture, l’Alimentation et l’Environnement – France

Résumé

The world is reaching a point where effectiveness of antibiotics could be completely compromised in the near future, if antimicrobial resistance (AMR) continues to spread globally. The intestinal microbiota of human and domestic animals is suspected to be the main reservoir of AMR organisms and therefore of antibiotic resistance genes (ARGs). Advances in DNA sequencing technology and more specially metagenomic approaches have shown that several ARGs are prevalent and shared between most gut microbiota. Although nucleotide diversity is documented for many ARGs, each variant following independent spread, the distribution of these variants in gut resistomes is still completely unknown. This study aims to understand if ARGs described as identical in different microbiomes are actually the same variants.

We adapted DESMAN, an algorithm developed to reconstruct strains from metagenomic data, to characterize different variants from a pool of ARGs. The tool was applied to ARGs detected in metagenomic cecal samples from chicken raised under different conditions. Among 22 genes analyzed, 15 have at least 2 stable variants, 7 (*ant(6)*, *blaTEM-1*, *erm(B)*, *erm(F)*, *tet(Q)*, *tet(L)* and *tet(32)*) showing variants with important differences in proportions between samples depending on rearing condition and age of animals. Sequence comparison between variants reconstructed by DESMAN and those identified from bacterial isolates revealed a perfect match. These results attest the reliability of the tool to reconstruct different variants of the same ARG directly from metagenomic data and to infer their relative proportion in different samples. It opens the way for further analysis to understand their relative abundance, persistence and transmission between microbiomes.

Mots-Clés: Antimicrobial resistance, gene variants, gut microbiota, Gibbs sampling method

*Intervenant

Characterization of the functional composition of the Human Gut Microbiome in Liver Cirrhosis and Colorectal Cancer (CRC) and identification of candidate biomarkers using AI

Baptiste HENNECART*¹, Sandy Frank KWAMOU NGAHA*¹, Florian PLAZA OÑATE¹, Vadim PULLER¹, Thomas MONCION¹, Edi PRIFTI^{2,3} and Raynald DE LAHONDÈS¹

¹ GMT Science, 75 route de Lyons-la-Forêt, 76000 Rouen, France

² IRD, Sorbonne Université, Unité de Modélisation Mathématique et Informatique des Systèmes Complexes, UMMISCO, 93143 Bondy, France

³ Sorbonne Université, INSERM, Nutrition and Obesities; Systemic approaches, NutriOmics, AP-HP, Hôpital Pitié-Salpêtrière, 75013 Paris, France

Corresponding Author: baptiste.hennecart@gmt.bio

Liver cirrhosis and colorectal cancer (CRC) are two pathologies associated with strong modifications in species composition of the human gut microbiome[1]–[3]. Differential Abundance Analysis (DAA) is typically used to characterize these changes. Changes in species abundance are usually the only biomarkers sought in this type of study. We have used several AI methods to uncover specific signature of these pathologies based upon these differences of abundance with success. However, there are several changes in the microbiome composition that cannot be detected by this approach (notably within a given species, strains may display different behaviors, switching from commensal to pathogens, due to limited differences in genetic content). Here we have re-analyzed different metagenomics studies using predictive functions quantification with a novel framework. We compared the genetic functional level approach with the state of art approach which quantifies abundance at the species level approach and how genic function level approach could enhance the performance of microbiome-based diagnosis.

It precisely aimed to investigate the metabolic functions of gut microbiomes in two distinct cohorts: liver cirrhosis[4] and colorectal cancer. Leveraging gene-level analysis of sequencing data, we employed KEGG orthology (KO)[5] data to quantify these functions. To accomplish this, we performed gene quantification using shotgun sequencing data produced from fecal samples in the aforementioned patient cohorts. The Integrated Gene Catalog 2 (IGC2)[6] served as a valuable resource in our analyses, further complemented by functional annotations obtained from the KEGG database, encompassing KO assignments, and from which we derived KEGG modules assignments.

Acknowledgements

The computational part of this study was sponsored by OVH, through the Startup Program.

References

- [1] R. Li, Z. Mao, X. Ye, and T. Zuo, ‘Human Gut Microbiome and Liver Diseases: From Correlation to Causation’, *Microorganisms*, vol. 9, no. 5, p. 1017, May 2021, doi: 10.3390/microorganisms9051017.
- [2] S. Singh *et al.*, ‘Implication of Obesity and Gut Microbiome Dysbiosis in the Etiology of Colorectal Cancer’, *Cancers*, vol. 15, no. 6, p. 1913, Mar. 2023, doi: 10.3390/cancers15061913.
- [3] J. Li, A. Zhang, F. Wu, and X. Wang, ‘Alterations in the Gut Microbiota and Their Metabolites in Colorectal Cancer: Recent Progress and Future Prospects’, *Front. Oncol.*, vol. 12, p. 841552, Feb. 2022, doi: 10.3389/fonc.2022.841552.
- [4] N. Qin *et al.*, ‘Alterations of the human gut microbiome in liver cirrhosis’, *Nature*, vol. 513, no. 7516, pp. 59–64, Sep. 2014, doi: 10.1038/nature13568.
- [5] M. Kanehisa, Y. Sato, M. Kawashima, M. Furumichi, and M. Tanabe, ‘KEGG as a reference resource for gene and protein annotation’, *Nucleic Acids Res.*, vol. 44, no. D1, pp. D457–D462, Jan. 2016, doi: 10.1093/nar/gkv1070.
- [6] C. Wen *et al.*, ‘Quantitative metagenomics reveals unique gut microbiome biomarkers in ankylosing spondylitis’, *Genome Biol.*, vol. 18, no. 1, p. 142, Dec. 2017, doi: 10.1186/s13059-017-1271-6.

Deciphering and cataloging the genomic and functional diversity of the French cheese microbiota

Hélène GARDON¹, Françoise IRLINGER¹, Céline DELBÈS², Éric DUGAT-BONY¹, Julia GENDRE¹, Olivier RUÉ³, Corinne CRUAUD⁵, Sébastien THEIL², Cécile CALLON², Valentin LOUX³, Pierre RENAULT⁴, Mahendra MARIADASSOU³, Frédéric GAVORY⁵, Valérie BARBE⁵, Cécile NEUVÉGLISE⁶, Vincent HERVÉ¹

¹ Université Paris-Saclay, INRAE, AgroParisTech, UMR SayFood, 91120, Palaiseau, France

² Université Clermont Auvergne, INRAE, VetAgro Sup, UMR 0545 Fromage, 20 Côte de Reyne, 15000 Aurillac, France

³ Université Paris-Saclay, INRAE, Unité MaIAGE, Domaine de Vilvert, 78352 Jouy-en-Josas, France

⁴ Université Paris-Saclay, INRAE, AgroParisTech, Micalis Institute, Domaine de Vilvert, 78352 Jouy-en-Josas, France

⁵ Génomique Métabolique, Genoscope, Institut de Biologie François Jacob, CEA, CNRS, Université Evry, Université Paris-Saclay, 91057 Evry, France

⁶ INRAE, Institut Agro, SPO, University Montpellier, 2 place Viala, 34060 Montpellier

Corresponding Authors: helene.gardon@inrae.fr, vincent.herve@inrae.fr

Cheeses are fermented products resulting from the complex interaction between the milk, the environment and the micro-organisms that compose them. The species found in this ecosystem originate from commercial starters, which are involved in the manufacturing and maturing processes, but also from endogenous microbial communities, which are influenced by the different cheese production practices. Although these particular communities are essential, little is known about their taxonomic and functional diversity, as well as the genetic traits of their adaptation to the cheese ecosystem. As such, the MétaPDOcheese project – "Grand Projet de Séquençage" France Génomique (2017 – 2023) – was initiated to explore the diversity of microbial communities inhabiting the Protected Designation of Origin (PDO) cheeses at the scale of the French territory.

By analysing 146 metagenomes from three environments – *i.e.*, milk, cheese core and rind – we review the bacterial, fungal and viral diversity of 44 PDO cheeses. We show that the *alpha* and *beta* diversity are influenced by the environment, the technology used (soft, hard or semi-hard) and the type of rind (bloomy, washed or natural), with a higher abundance of *i)* bacteria in hard and semi-hard cheeses and *ii)* eukaryotes in bloomy rind and blue mold cheeses. This taxonomic profiling also highlights a positive relationship between the abundance of bacterial and viral sequences, suggesting a prevalence of bacteriophages in hard and semi-hard cheeses in particular. Moreover, the functional analysis reveals a dissimilarity between the three environments, but also between technologies/type of rind.

Based on the analysis of co-occurring species within these microbial communities, some of which are potential keystone species of the ecosystem, a set of 373 genomes of bacterial strains isolated from cheeses were sequenced. In addition to this culturomic approach, we reconstructed 1119 metagenome-assembled genomes (MAGs) of good quality from the 146 metagenomes. By comparison with the reference sequences currently available in public databases, the analysis of all these genomes revealed a set of 259 previously not described, and therefore potentially new, species belonging to genera of interest in the cheese ecosystem (*e.g.*, *Psychrobacter*, *Halomonas*, *Brevibacterium*). Altogether, combining culturomic and metagenomic approaches has given us an access to an uncovered diversity of non-inoculated micro-organisms.

In addition, from genomic and metagenomic data, we have built a catalogue of functionally and taxonomically annotated reference proteins, that includes more than 15 millions non-redundant proteins and covered the metabolic diversity of the cheese ecosystem and the 44 French PDO cheeses. The construction of this protein catalogue constitutes an accessible tool for future understanding of the structure and function of microbial communities in other cheese ecosystems.

Identification of a microbiome, the advantages of the metagenomic method over the classical 16S method

Aurélie PETICCA¹, Christophe KLOPP², Mostefa FODIL¹, Benoît CHENAIS¹, Nathalie CASSE¹ and François SABOT³

¹ BIOSSE, University of Le Mans, F-72085 Le Mans, France

² Genotoul Bioinfo, BioInfoMics, MIAT UR875, INRAE, F-31326 Castanet-Tolosan, France

³ DIADE, University of Montpellier, CIRAD, IRD, F-34394 Montpellier, France

Corresponding Author: aurelie.peticca@univ-lemans.fr

Abstract

Haslea ostrearia is a marine microalgae producing a blue pigment of great interest. Firstly, from an economic point of view, this marenine pigment colors oyster gills green which increases their market value [1]. Secondly, marenine could also have biomedical applications, such as an antitumoral effect against lung/kidney carcinoma and melanoma cell lines [2, 3]. However, *H. ostrearia* is poorly known at the genetic level. Any attempt to assemble its genome is complicated by the large number of bacteria living with it [4]. Their presence seems to be necessary for the microalgae survival. It therefore became necessary to identify these bacteria to improve culture protocols by trying to reduce their presence in the sequencing data. This decrease in bacteria concentration should then allow us a better access to the microalgae genome.

To this end, we had two solutions. The first "classic" one, consisted in sequencing a 16S rRNA gene hypervariable region and performing a metataxonomic analysis using a reference 16S rRNA database. The second was to use whole genome sequencing data to assemble the metagenome and extract taxonomic information from it (metagenomic strategy). In order to ensure robust results, we used both methods to list bacteria present in *H. ostrearia* cultures. Comparing the results allowed us to understand the impact of the chosen method and to see the advantage of the metagenomic strategy. For the 16S method, a known bacterial community was used to compare the results between two 16S regions V1-V3 and V3-V4. The results were highly variable between regions. V1-V3 even failed to identify an entire class of bacteria: *Gammaproteobacteria*. The bacteria identified in our cultures were also less specific than those found through whole genome sequencing. By recovering entire 16S genes from the metagenome assembly, the identification step was more accurate than using only a small proportion of it. Not only did we find the same bacteria that were highly present in the 16S sequencing, but we also identified some of them to species level. In addition, the metagenomic strategy enabled the assembly of completed or almost completed genomes. Collected genetic information was therefore much larger with this method enabling other analysis at a later stage. For example, it permitted to search for antibiotic or microbiotic resistance genes in the assemblies. This could help to understand the biological processes occurring in our cultures and facilitate cultivation protocol improvement.

Acknowledgements

This work was supported by the LANCOM project (Pays de la Loire region) and by Le Mans Metropole. Many thanks for the Genotoul GeT-Place and GeT-Biopusces platforms for all the sequencing data.

References

1. F. Piveteau, 'Aroma of oyster *Crassostrea gigas*: effect of supplementation with the microalgae *Skeletonema costatum*', Doctoral dissertation, Nantes University, 1999. Accessed: Dec. 15, 2022. [Online]. Available: <https://www.theses.fr/1999NANT2017>
2. S. Méresse et al., '*Haslea ostrearia* Pigment Marenine Affects Key Actors of Neuroinflammation and Decreases Cell Migration in Murine Neuroglial Cell Model', *IJMS*, vol. 24, no. 6, p. 5388, Mar. 2023, doi: 10.3390/ijms24065388
3. D. Carbonnelle et al., 'Antitumor and antiproliferative effects of an aqueous extract from the marine diatom *Haslea ostrearia* (Simonsen) against solid tumors: lung carcinoma (NSCLC-N6), kidney carcinoma (E39) and melanoma (M96) cell lines', *Anticancer Res*, vol. 19, no. 1A, pp. 621–624, 1999
4. A. Peticca et al. 'The need to stabilize the bacterial community present in the cultures of the marine diatom *Haslea ostrearia*', In Review, preprint, Apr. 2023. doi: 10.21203/rs.3.rs-2751923/v1.

Microbial diversity of *Beggiatoa* mats reveal a different taxonomic profile from marine mangrove sediments in urban and natural sites of the Caribbean.

Mariana MARTÍNEZ-NORIEGA^{1*}, Patrick JEAN-LUIS² and Fidel Alejandro SÁNCHEZ-FLORES¹, Silvina GONZALEZ-RIZZO^{2*}

¹ Institute of Biotechnology (UNAM), Av. Universidad 2001, 62210, Cuernavaca, Mexico

² Institut de Systématique, Evolution, Biodiversité (ISYEB), MNHN, CRNS, Sorbonne Université EPHE, Université des Antilles, 97110 Campus Fouillole, Point-Pitre, Guadeloupe

* Corresponding author: silvina.gonzalez-rizzo@univ-antilles.fr, mariana.martinez@ibt.unam.mx

Mangroves ecosystems cover 60-75% of the world's tropical and subtropical coastline.

They are among the most productive marine ecosystems on the planet. They harbor a rich diversity of animal and microorganisms and offer important ecological services. They are characterized by a high turnover of organic matter mediated by microbial processes [1]. Regarding microbial communities the studies have mainly focused on the microbial communities of the terrestrial mangrove's environments (i.e. mangrove soils and sediments) [2,3], and little is known concerning the marine microbial communities of this ecosystem [4,5].

Microbial communities that grow as microbial mats are large filamentous microorganisms that form an entangled mass with the submerged sediment particles on which they grow [6]. Microbial mats are a widespread phenomenon and are found from deep to coastal environments, and from cold to tropical waters [6,7].

These mats are considered an ecological niche and play an important role in the food web of their environment, by ensuring the production of primary organic matter necessary for the development of many other organisms [7,8].

Previous work has identified the two major colorless sulfur bacterial forming white's mats living on the surface of marine mangrove sediments in Guadeloupe island [9]. However, the whole microbial community's diversity that constitute these mats have not been studied nor the contribution of these communities to the biogeochemical cycles nor their resilience in response to the urban pressure conditions of this marine ecosystem.

Metabarcoding analyses supported by high throughput sequencing provide a method to evaluate the microbial community in terms of both taxonomy and potential functioning.

Thus, it was therefore used to better understand the functioning of microbial mats in the coastal ecosystems of Guadeloupe and the impact of urban pressure on this ecosystem through comparative analyses of microbial communities diversity of *Beggiatoa* mats found in natural, protected and urban marine mangroves areas along to *la Rivière Salée* in Guadeloupe island (FWI).

Using 44 samples from 11 sampling sites, we investigated the biogeochemical properties as well as the microbial diversity of *Beggiatoa* mats formed in marine mangroves sediments. Preliminary analyses regarding the biodiversity rRNA gene metabarcoding datasets revealed a distinct taxonomic profiling of microbial communities of *Beggiatoa* mats between natural and urban marine mangrove sediments. However, the physicochemical properties measured have not shown any strong influence over the taxonomical profile variations.

Altogether, the data generated by this project will contribute to better understand the microbial community assembly and microbial processes in marine mangrove sediments at the scale of a Caribbean Island.

References

1. Gina Holguin, Patricia Vazquez, and Yoav Bashan. "The role of sediment microorganisms in the productivity, conservation, and rehabilitation of mangrove ecosystems: an overview." *Biology and fertility of soils* 33, 265-278, 2001
2. Janaina Rigonato, Angela D. Kent, Danillo O. Alvarenga, Fernando D. Andreote, Raphael M. Beirigo, Pablo Vidal-Torrado, and Marli F. Fiore. "Drivers of cyanobacterial diversity and community composition in mangrove soils in south-east Brazil." *Environmental Microbiology* 15, no. 4, 1103-1114, 2013

3. Maryeimy Varon-Lopez, Armando Cavalcante Franco Dias, Cristiane Cipolla Fasanella, Ademir Durrer, Itamar Soares Melo, Eiko Eurya Kuramae, and Fernando Dini Andreote. "Sulphur-oxidizing and sulphate-reducing communities in Brazilian mangrove sediments." *Environmental Microbiology* 16, no. 3, 845-855, 2014
4. Shamina M, Saranya T, Ram AT. Cyanobacterial biodiversity at mangrove vegetation of Kadalundi, Kerala. *J Microbiol* 3, 15–16, 2014
5. Sarah M. Allard, Matthew T. Costa, Ashley N. Bulseco, Véronique Helfer, Laetitia GE Wilkins, Christiane Hassenrück, Karsten Zengler et al. "Introducing the mangrove microbiome initiative: identifying microbial research priorities and approaches to better understand, protect, and rehabilitate mangrove ecosystems." *MSystems* 5, no. 5, e00658-20, 2020
6. Tom Fenchel, and Catherine Bernard. "Mats of colourless sulphur bacteria. I. Major microbial processes." *Marine Ecology Progress Series* 128, 161-170, 1995
7. Jody W. Deming, Anna-Louise Reysenbach, Stephen A. Macko, and Craig R. Smith. "Evidence for the microbial basis of a chemoautotrophic invertebrate community at a whale fall on the deep seafloor: Bone-colonizing bacteria and invertebrate endosymbionts." *Microscopy research and technique* 37, no. 2, 162-170, 1997
8. Lucas J. Stal, "Cyanobacterial mats and stromatolites." *Ecology of cyanobacteria II: their diversity in space and time*, 65-125, 2012
9. Maïtena Jean RN, Silvina Gonzalez-Rizzo, Pauline Gauffre-Autelin, Sabine K. Lengger, Stefan Schouten, and Olivier Gros. "Two new *Beggiatoa* species inhabiting marine mangrove sediments in the Caribbean." *PloS one* 10, no. 2, e0117832, 2015

Suitability of Nanopore adaptative sampling for metabarcoding approaches

Is it possible to directly remove chloroplast sequences from algal samples during sequencing?

Coralie Rousseau¹, Erwan Legeay², Gwenn Tanguy², Yacine Badis¹, Catherine Leblanc¹, Simon Dittami¹

¹ Integrative Biology of Marine Models (LBIM), UMR8227, Sorbonne Université, CNRS, Station Biologique de Roscoff, France
² Genomer Plateform, FR2424, Sorbonne Université, CNRS, Station Biologique de Roscoff, 29680 Roscoff, France

Background

The description of epi- and endobacterial communities through short-read metabarcoding analyses is the most popular way to identify a large proportion of non-cultivable species. However, this analysis is based on 16S rRNA amplification which also amplifies plastid DNA of the host. Without specific primers design to deplete plastid reads, the majority of reads will belong to them and for long-read metabarcoding analyses, such primers are not yet available.

Recently, "adaptative sampling", a method of software-controlled during sequencing has been developed for the Oxford Nanopore platform. It is possible to enrich or eject specialty reads in real-time during sequencing^{1,2}. This method have already been tested in metagenomic analysis and have shown promising results^{3,4}. Here we have tested the suitability of adaptative sampling in metabarcoding analyses to remove plastid reads.

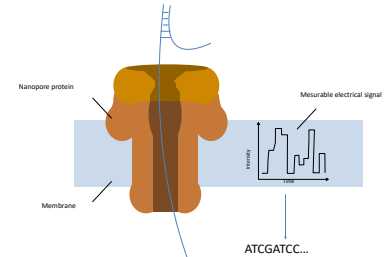


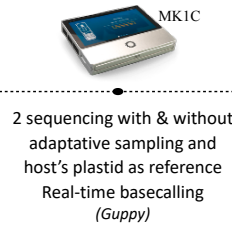
Figure 1. Principle of Nanopore sequencing. DNA passes through Nanopore protein, generates measurable electrical signal which is converted to nucleotide sequence.

Method



24 DNA extractions of *Ascophyllum nodosum* (November & March)

Full 16S rRNA amplicon (27F-1492R) Nanopore's library preparation



2 sequencing with & without adaptative sampling and host's plastid as reference Real-time basecalling (*Guppy*)

Control quality (*Fastqc*)

Assign taxonomy (*Kraken2* and *SILVA* database)

Results

Table 1. Sequencing results of the two runs.

	Without adaptative sampling	With adaptative sampling
Number total reads	2,524,217	2,491,252
Mean length	1296	755
Mean Q20 (%)	43	57
Total abundance of chloroplast reads	1,674,221	1,629,764

Both sequencing runs generated similar numbers of reads. We have not enrich the run in bacterial reads by depleting plastid DNA. The primary difference between the runs was the mean read length, which was shorter in the run with adaptive sampling.

The histograms in Figure 2 show that virtually all of the plastid reads were truncated at approximately 700 bp in the adaptative sampling run, but also a large proportion of the bacterial reads.

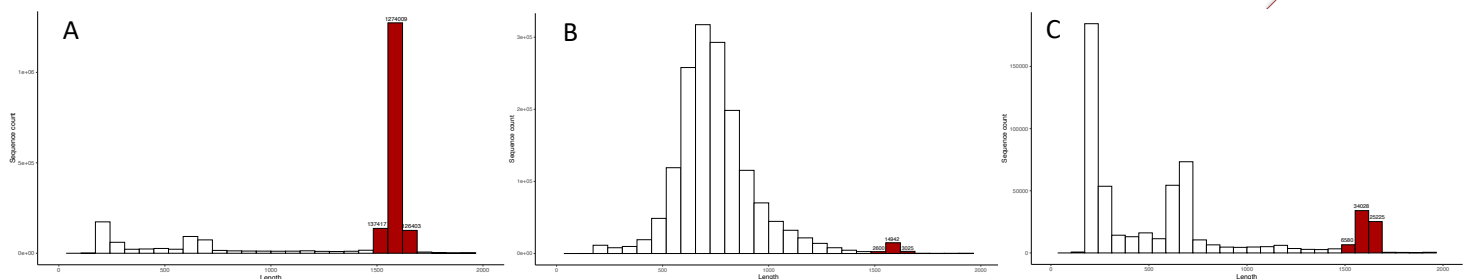


Figure 2. What happened during the two sequencing? A) First run - Without adaptative sampling B) Second run - With adaptative sampling. Only plastid sequences are shown C) Second run - With adaptative sampling. Only bacteria sequences are shown. Red color represents the expected size for bacteria (e.g. 1500-1700 pb).

Conclusion

Based on our experiment, we conclude that adaptative sampling with the MK1C is not suitable for the removal of plastid sequences in 16S metabarcoding analyses.

Two important limitations:

1. The great similarity between plastid and bacterial reads leads to the rejection of a large proportion of bacterial reads.
2. The default algorithm implemented on the MK1C is not adjustable

Alternative base-calling and alignment methods as well as the optimization of alignment parameters (percent identity for alignment against reference, the number of bases align to reference) could lead to improvements.

¹ Loose, Matthew, Sunir Malla, et Michael Stout. 2016. « Real-Time Selective Sequencing Using Nanopore Technology ». *Nature Methods* 13(2): 151-54.

² Payne, Alexander, Nadine Holmes, Thomas Clarke, Rory Munro, Bisrat J. Debebe, et Matthew Loose. 2021. « Redfish Enables Targeted Nanopore Sequencing of Gigabase-Sized Genomes ». *Nature Biotechnology* 39(4): 442-50.

³ Payne, Alexander, Nadine Holmes, Thomas Clarke, Rory Munro, Bisrat Debebe, et Matthew Loose. 2020. « Nanopore Adaptive Sequencing for Mixed Samples, Whole Exome Capture and Targeted Panels ».

⁴ Martin, Samuel, Darren Heavens, Yuxuan Lan, Samuel Horsfield, Matthew D. Clark, et Richard M. Leggett. 2022. « Nanopore Adaptive Sampling: A Tool for Enrichment of Low Abundance Species in Metagenomic Samples ». *Genome Biology* 23(1): 11.

Acknowledgements: We thank Germain Cheignon for his advice

Tracing human-borne bacterial contaminants from wastewater treatment plants to coral reefs in the Caribbean Sea

Ander Urrutia^{1,2}, PascalJean Lopez^{2,3}, Josie Lambourdiere^{2,3}, Fred Burner^{1,2}, Sébastien Cordonnier^{1,2}, Isabelle Nasso⁴, Mélissa Bocaly⁴, Malika ReneTrouillefou^{1,2}

¹Université des Antilles(UA),Laboratoire Biologie Marine, 97157, Pointe-à-Pitre, France

²Laboratoire de Biologie des Organismes et des Ecosystèmes Aquatiques (MNHN, CNRS, SU, IRD, UCN, UA), 43 rue Cuvier, 75231, Paris, France

³Centre National de la Recherche Scientifique (CNRS),3 rue Michel-Ange, 75794, Paris, France

⁴Office de l'eau de Martinique (ODE), 7 avenue Condorcet, 97201, Fort-de-France,France

Corresponding Author: firstname.lastname@email.fr

For the last 20 years, the oceanic surface occupied by living coral reefs has been diminishing throughout the French Antilles. Concomitantly, the number of locations rated as “insufficient” in regard to the water quality for human health, has been noticeably increasing in these oversea territories. Since the water-quality rating is largely based on the presence and abundance of faecal bacteria, most significantly *Escherichia coli*, *Enterococcus* sp., and *Streptococcus* sp., we hypothesized this and other prokaryotes to have a role in linking both, coral and water quality decrease. In this context, our project was set up with two main objectives. In the one hand, conducting a review of potential sources of such bacteria in the Caribbean Sea, paying special attention to geographical and seasonal variations. On the other hand, to identify and compare the bacterial communities associated to coral reefs in environments differentially exposed to anthropogenic sources. Our preliminary results, which graphicate official data (IFRECOR, 2020; Ministère de la Santé, 2023), show the water quality and living coral reef extension of the Guadeloupe-an archipelago to be deteriorating particularly quick. A situation that does not necessarily mirror the patterns observed in other islands of the French Antilles, at least during the last few years. With wider-scale factors seemingly unable to explain, at least on its own, this regional variation, our approach in the TRACMIC project ‘*TRAçage des Contaminants MICrobiens d’origine humaine en provenance des stations d’assainissement dans les Antilles Françaises*’ is to analyze the local variation of human-borne bacterial communities in waste water treatment plants. Given the more static and long-term nature of the flux derived from them, WWTPs were selected in collaboration with the OFB to conduct an eDNA-based analysis. However, we acknowledge that they do not represent the only source of faecal bacteria. As part of this first approach, water from the intake and the discharge of 19 WWTPs was sampled using a continuous (24 h) sampler; Basse Terre (n = 9), Grande Terre (n = 5), Marie Galante (n = 1), and Martinique (n = 4). Additionally, water from the recipient rivers and coastal locations, including areas of low and high living-coral surface, was also recovered for analysis. In total, 125 eDNA samples (250 replicates) were filtered using 0.2 µm filters and the Millipore vacum system, with final volumes ranging between 10 and 120 ml. Samples are, at the time of writing, being amplified (and Illumina sequenced (Miseq: 2 x 300 bp) before trimming, merging, de-replicating, de-noising, chimera-filtering, and read-clustering into OTUs. In this fashion, we aim to compare the bacterial communities from the discharge areas (fresh- and sea-water) with those associated to coral reefs. By linking community composition with distance to the WWTPs ejection site, the intention is to identify new faecal-pollution biomarkers especially orientated to coral health. While the fulfillment of Koch’s postulates is rare and complex for coral-infecting diseases, several bacterial taxa have been already associated to coral loss in the literature. Thus, our work will also aim to find such specific genera and species in our database. Currently, we are identifying some of those taxa of interest such as *Vibrio coralliilyticus*, *Pseudoscillatoria coralii*, or *Desulfovibrio* sp. in waters of the French Antilles, by blasting and phylogenetically comparing (Maximum Likelihood & Bayesian inference) their 16S rRNA against existing databases from already published eDNA samplings. Future steps will include the quantification of such biomarkers via qPCR from the coral reefs to the source, as well as a closer look at the water-coral interface to examine the type of association between bacteria and coral.

A global catalogue of genomes and protein sequences from the termite microbiome

Nachida TADRENT¹, Franck DEDEINE¹ and Vincent HERVE^{1,2}

¹ Institut de Recherche sur la Biologie de l'Insecte, UMR 7261, CNRS-Université de Tours, 37200, Tours, France

² Université Paris-Saclay, INRAE, AgroParisTech, UMR SayFood, 91120, Palaiseau, France

Corresponding Authors: nachida.tadrent@univ-tours.fr, vincent.herve@inrae.fr

Advances in whole metagenome shotgun sequencing and associated bioinformatics approaches facilitate environmental studies of microorganisms at both the community and genome levels. Several efforts have been made to centralize information and make it more accessible to microbial ecologists by establishing genome and/or protein catalogues specific to a given environment (e.g., the genomic catalogue of earth's microbiomes, the gene catalogue of the mouse gut, the catalogue of reference genomes from the human gut microbiome, the ocean gene atlas). These valuable resources allow a better understanding of the structure, dynamics and functions of microbial communities. In this study, we are interested in the microbiota of termites, which are the most efficient insects to degrade lignocellulose, and are therefore ecologically and economically important. They possess a complex and diverse gut microbial community (Bacteria, Archaea, Eukaryota) that contributes greatly to the lignocellulosic digestion process. Although many studies have described the taxonomic diversity and community structure of the termite microbiota, the roles of individual lineages involved in the digestive and metabolic functions remain poorly understood.

To reveal the functional and metabolic potential of the microbial community, we built the largest resource of microbial genomes associated to termites as well as the related catalogue of proteins. First, a literature search of metagenome and genome collections of termite hosts and all their associated microbes (Bacteria, Archaea, Protists, Fungi) was performed according to the PRISMA [1] protocol. Second, using *SnakeMAGs* [2], we reconstructed 514 bacterial and archaeal metagenome-assembled genomes (MAGs) with completeness $\geq 50\%$ and contamination $< 10\%$ (according to CheckM [3]) and not chimeric (according to GUNC [4]) from 80 publicly available Illumina shotgun metagenomes. The use of metagenomics allowed us to overcome the culture-dependant biases and thus to cover greater diversity. This dataset was further enriched by 976 prokaryotic and 28 eukaryotic genomes (from both MAGs and microbial isolates) retrieved from public databases. Altogether we constructed a comprehensive catalogue containing 1518 genomes. In addition, we conducted a protein-level assembly using Plasm [5] directly from cleaned metagenomic reads to massively recover microbial proteins associated with termites. Predicted proteins from collected genomes were also added to the Plasm results and clustered using MMseq2 [6] to produce a large catalogue composed of millions of non-redundant proteins.

This data set, which integrates the largest collection of termite-related microbial genomes and its associated protein catalogues, will facilitate genome-centric studies and thus will improve our understanding of these fascinating host-microbiota interactions and their evolutionary dynamics.

References

1. Page, M. J. et al. The PRISMA 2020 statement: An updated guideline for reporting systematic reviews. In *The BMJ* (Vol. 372). (2021).
2. Tadrent, N., Dedeine, F., & Hervé, V. *SnakeMAGs*: a simple, efficient, flexible and scalable workflow to reconstruct prokaryotic genomes from metagenomes [version 2; peer review: 2 approved]. *F1000Research*, 11, 1522. (2023). <https://doi.org/10.12688/f1000research.128091.2>
3. Parks, D. H. et al. CheckM: Assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes. *Genome Research*, 25(7), 1043–1055. (2015).
4. Orakov, A. et al. GUNC: detection of chimerism and contamination in prokaryotic genomes. *Genome Biology*, 22(1). (2021).
5. Steinegger, M., Mirdita, M., & Söding, J. Protein-level assembly increases protein sequence recovery from metagenomic samples manyfold. *Nature Methods* 2019 16:7, 16(7), 603–606. (2019).
6. Steinegger, M., & Söding, J. Clustering huge protein sequence sets in linear time. *Nature Communications*, 9(1). (2018).

ABRomics: An integrated multi-omics platform for antibiotic resistance research and public health

Julie Lao^{1, 3}, *Pierre Marin*^{1, 2}, *Romain Dallet*^{1, 4}, *Fabien Mareuil*⁷, *Alix De Toisy*, *Kenzo-Hugo Hillion*⁵, *Aurélien Birer*⁶, *Nadia Goué*², *Richard Bonnet*⁶, *Etienne Ruppé*³, *Gildas Le Corguillé*^{1, 4}, *Philippe Glaser*⁸, *Claudine Médigue*^{1, 9}

¹CNRS, Institut Français de Bioinformatique, IFB-core, UAR 3601 - Évry (France), ²AuBi platform, Mésocentre, Université Clermont-Auvergne - Aubière (France), ³Université Paris Cité and Université Sorbonne Paris Nord, Inserm, IAME - Paris (France), ⁴Sorbonne Université, CNRS, FR2424, ABiMS, Station Biologique - Roscoff (France), ⁵Faculté des Sciences Techniques, Université de Nantes, F- 44000 - Nantes (France), ⁶CHRU Clermont-Ferrand Gabriel-Montpied - Clermont-Ferrand (France), ⁷Institut Pasteur, Université Paris Cité, Bioinformatics and Biostatistics Hub, F-75015 - Paris (France), ⁸Institut Pasteur, Unité EERA, CNRS UMR604, F-75015 - Paris (France), ⁹CNRS UMR8030, Université Evry-Val-d'Essonne, CEA, Genoscope, LABGeM - Évry (France)

Antibiotic resistance (ABR) is a major public health issue prioritized for mitigation by international institutions. Multidrug resistant bacteria (MDRB) and Antibiotic Resistance Genes (ARGs) carried by mobile genetic elements spread between the human, animal, and environmental sectors. Whole Genome Sequencing (WGS) is used for molecular typing purposes at the highest resolution. It provides identification of ARGs and their genetic supports as well as mutations leading to a decrease in antibiotic susceptibility. Epidemiological and WGS data are used for tracking MDRB in hospital outbreaks but also across the animal and environmental sectors. Sharing and interoperability of high-quality data (sequence and metadata) are key requirements for addressing the spatio-temporal dissemination of MDRB. To this aim, the French Priority Plan on ABR has funded the development of an online, open platform dedicated to antibiotic resistance.

We are establishing a repository of structured, interoperable, standardized, and well-annotated multi-omics data with tailored mathematical and bioinformatics tools to answer generic and specific research questions related to ABR. The ABRomics platform includes standardized pipelines to run ABR analyses of WGS from pathogenic strains supported with integrated databases (ARG, sequence types [ST], virulence factors [VF]). Uploading data, launching pipelines, viewing and cross-referencing enriched results will be achieved through easy-to-use web interfaces. ABRomics β -version integrating the ABR detection genomic pipeline and other markers such as ST, and VF will be available to the consortium in summer 2023 and to the whole microbial research community by the end of 2023. Core-genome multi-locus sequence typing, relationships between strains and metagenomics pipelines will next be made available.

Acknowledgement of grants and fundings:

This work is financially supported by the French Priority Research Programme on Antimicrobial Resistance (PPR antibioresistance), coordinated by Inserm and funded by the Secrétaire General Pour L'investissement (SGFI).



Khaoula Ziane
CEA / Genoscope

Résumé du poster :

Titre : Pipeline d'annotation de génomes à partir de données de séquence d'ARN messagers

Les technologies de séquençage actuelles permettent de produire à grande échelle des génomes de haute qualité pour un nombre croissant d'organismes vivants très divers. Pour annoter ces génomes eucaryotes, il faut disposer de méthodes rapides et exhaustives et facilement adaptables aux organismes étudiés. Pour une annotation exhaustive, il est important de disposer de séquences transcriptomiques issues du séquençage d'ARN messagers (RNA-Seq). L'alignement de ces données sur le génome permet de mettre en évidence les régions transcrites d'une séquence et donc améliorer la prédiction de gènes.

Ici, nous détaillerons un pipeline automatisé que nous avons développé, se basant sur ces données RNA-Seq pour annoter les génomes. Notre méthode peut utiliser des données RNA-Seq fournies par l'utilisateur ou extraire directement ces données depuis les banques de données publiques. La méthode proposée permet d'obtenir des résultats précis pour l'annotation de génomes eucaryotes. Cette approche est applicable à plusieurs génomes en parallèle, offrant ainsi une solution rapide et efficace pour répondre à la demande croissante en matière d'annotation de génomes. Nous détaillerons également la mise en œuvre de ce pipeline à l'aide de plusieurs génomes, démontrant ainsi son potentiel en termes de traitement de données.

Bioinformatics challenges in the analysis of gastruloid time-series single-cell RNA-seq data

Céline CHEVALIER¹, Laurent ARGIRO¹, Caroline CHOQUET¹, Nitya NANDKISHORE¹, Adeline GHATA¹, Stéphane ZAFFRAN¹, Fabienne LESCROART¹ and Anaïs BAUDOT^{1,2}

¹ Aix Marseille Univ, INSERM, MMG, 13385, Marseille, France
² Barcelona Supercomputing Center (BSC), 08034, Barcelona, Spain

Corresponding author: `celine.chevalier@univ-amu.fr`

1 Biological context

We used gastruloids to investigate the formation of different cell types and heart-associated tissues during development. Gastruloids are in vitro models, which allow for controlled and reproducible experiments. We applied a gastruloid culture protocol to model cardiopharyngeal mesoderm specification into cardiac and muscle tissues (Laurent Argiro et al. in preparation). We performed scRNA-seq experiments on day 4, day 5, day 6, and day 11 of gastruloid development. The objective is to assess the ability of gastruloids to mimic cardiogenesis and myogenesis (Laurent Argiro et al. in preparation).

2 scRNA-seq data analysis

We first implemented a standard pipeline for scRNA-seq analysis using the Seurat framework [1]. In quality control (QC), we paid attention to the biological material of the scRNA-seq samples to adapt the settings, as cells deemed low quality in one dataset may be valuable in another. Besides, doublets are described as a source of technical noise as they can lead to misinterpretation of the data. We used the DoubletFinder tool [2] to identify doublets in our datasets. But, some cells identified as doublets formed a cluster, and we had to consider the biological signal sent by these clustered doublets to ensure an accurate interpretation of the data. After normalization, scaling and Principal Components Analysis, the standard pipeline ends with cluster identification, UMAP visualisation and differential expression analysis.

3 Specific challenges of the time-series gastruloids scRNA-seq

We annotated the cell types using the reference atlas [3] by transfer learning. However, this reference atlas does not cover all our time points, leading to some challenges. In this poster, I will discuss how the reference atlas covers each of our datasets and highlight limitations in terms of cell type annotation.

For transcriptional trajectories reconstruction, we had to choose between merging or integrating the time-series scRNA-seq datasets. We tested the standard integration framework of Seurat [1] and Harmony [4]. In addition, we tried sequential integration of the datasets to better represent the time passing from the first time point dataset to the last one. We also faced batch biases. In this poster, I will discuss the results obtained by testing different tools and protocols.

Acknowledgements

We thank Lionel Spinelli and the CIML bioinformatics platform for expert advice on bioinformatics analyses. Computations were run on the Core Cluster of the Institut Français de Bioinformatique (IFB) (ANR-11-INBS-0013).

References

- [1] Yuhan Hao et al. Integrated analysis of multimodal single-cell data. *Cell*, (13):3573–3587.e29, 2021.
- [2] Christopher S. McGinnis, Lyndsay M. Murrow, and Zev J. Gartner. DoubletFinder: Doublet detection in single-cell RNA sequencing data using artificial nearest neighbors. *Cell Systems*, 8:329–337.e4, 2019.
- [3] Blanca Pijuan-Sala et al. A single-cell molecular map of mouse gastrulation and early organogenesis. *Nature*, (7745):490–495, 2019.
- [4] Ilya Korsunsky et al. Fast, sensitive and accurate integration of single-cell data with harmony. *nature methods*, (16):1289–1296, 2019.

Double approche, expérimentale et bio-informatique pour l'étude de l'épissage des ARN pré-messagers dans les cancers MSI

Fabien Kon-Sun-Tack^{1*}, Enora Le Scanf^{1*}, Laurent Corcos¹, Gaëlle Marenne¹

¹ Univ Brest, Inserm, EFS, UMR 1078, GGB, F-29200 Brest, France

* les auteurs ont contribué à part égale à ce travail

email:fabien.konsuntack@univ-brest.fr ; enora.lescanf@univ-brest.fr

Les cellules cancéreuses MSI (MicroSatellites Instables) possèdent un système de réparation de l'ADN nommé MMR (*MisMatch Repair*) défectueux, ce qui entraîne une instabilité des microsatellites, des répétitions nucléotidiques dispersées dans le génome. Un microsatellite particulier localisé en 3' des introns, nommé *polypyrimidine tract* (PyT), participe à la reconnaissance de l'exon en aval lors de l'épissage des ARN pré-messagers en fixant la protéine U2AF2. Cette protéine forme un hétérodimère avec la protéine U2AF1 qui reconnaît le dinucléotide AG (site 3' d'épissage). Une altération du PyT, notamment une délétion d'un ou de quelques nucléotides peut entraîner l'exclusion partielle de l'exon en aval dans l'ARN messager (*e.g.*, HSP110). En dépit de séquences consensus retrouvées dans la majorité des introns, l'épissage peut être variable, en particulier sous l'effet de mutations dans la région 3' des ARN pré-messagers. Des outils informatiques sont disponibles pour identifier les sites canoniques d'épissage de l'ARN, et se sont révélés performants [1]. Il a aussi été développé des outils capables de prédire l'effet de mutations dans la séquence du PyT, ou à son voisinage, sur l'épissage.

L'objectif de notre travail est d'évaluer ces outils dans le cadre de délétions dans le PyT. Nous nous sommes focalisés sur l'étude de délétions dans les PyT de 4 gènes : DNAJC18, KDM6A, PTP4A2 et TRAF3IP1 dont des mutations dans le PyT sont trouvées dans les cancers digestifs MSI.

Pour étudier l'impact de l'altération du PyT sur l'épissage, les ARN pré-messagers de lignées cellulaires MSI porteuses de mutations sont comparés aux ARN pré-messagers de lignées cellulaires MSI non mutées ou MSS (microsatellite stable). Nous avons testé deux outils bio-informatiques en accès libre : Human Splicing Finder [2] et CADD-Splice [3].

Les résultats expérimentaux montrent que la réduction de la taille du PyT entraîne une exclusion partielle de l'exon en aval dans l'ARN messager mature. HSF prédit une altération de la sélection du site d'épissage, notamment une activation d'un site accepteur cryptique. L'outil CADD-Splice classe les 4 évènements parmi les 10% les plus délétères au niveau du génome (score PHRED > 10).

Les outils bio-informatiques prédisent donc bien une altération de l'épissage, résultant de l'absence d'une ou de plusieurs pyrimidines. En revanche, la conséquence principale n'est pas en accord avec les observations expérimentales : les sites cryptiques d'épissage sont majoritairement prédits, plutôt que les sauts d'exons, qui n'utilisent pas de sites cryptiques, mais les sites consensus introniques. Il est envisagé que la position exacte de chacune des pyrimidines au sein du PyT ne soit pas prise en compte de la même façon (n'ait pas le même poids) par les algorithmes de prédiction, en particulier relativement à la position du point de branchement ou du site accepteur AG. Une analyse étendue à d'autres mutations dans un ensemble d'autres gènes pourrait permettre d'évaluer cette hypothèse plus en avant.

Références

1. Fernando Carazo, Juan P Romero, Angel Rubio, Upstream analysis of alternative splicing: a review of computational approaches to predict context-dependent splicing factors, *Briefings in Bioinformatics*, Volume 20, Issue 4, July 2019, Pages 1358–1375, <https://doi.org/10.1093/bib/bby005>
2. Desmet, François-Olivier et al. "Human Splicing Finder: an online bioinformatics tool to predict splicing signals." *Nucleic acids research* vol. 37,9 (2009): e67. doi:10.1093/nar/gkp215
3. Rentzsch, P., Schubach, M., Shendure, J. et al. CADD-Splice—improving genome-wide variant effect prediction using deep learning-derived splice scores. *Genome Med* 13, 31 (2021).

Evaluation of Oxford Nanopore R9.4.1/Kit10 , R10.4/Kit12 and R10.4.1/Kit14 sequencing for minority variants analysis

Théophile BOYER¹, Richard CHALVIGNAC¹, Bruno SIMON¹, Antonin BAL¹, Gregory DESTRAS¹
and Laurence JOSSET¹

¹Laboratoire de Virologie, Institut des Agents Infectieux, Laboratoire associé au Centre National de Référence des virus des infections respiratoires, Hospices Civils de Lyon, F-69004, Lyon, France

GenEPII sequencing platform, Institut des Agents Infectieux, Hospices Civils de Lyon, F-69004, Lyon, France

Corresponding Author: theophile.boyer@chu-lyon.fr

Oxford Nanopore technologies (ONT) have been used to rapidly produce accurate consensus-level sequence for SARS-CoV-2 genomic surveillance [1]. Because of elevated sequencing error rate, ONT have been poorly used to analyze within-host diversity of pathogens. Improvement in accuracy of ONT sequencing would allow better characterization of biological event including selection of resistance mutations, coinfection or recombination already observed with current NGS technologies [2,3].

In this study, we evaluated the three latest chemistries developed by ONT (Kit 10, Kit 12, Kit 14) compared with Illumina for minority variants analysis characterization. This analysis was performed on SARS-CoV-2 viral isolates, Delta (B.1.617.2) and Omicron (BA.1), mixed using different ratios [2] and sequenced using the Midnight amplicon protocol. Sequencing reads were analyzed using our in-house seqmet pipeline (<https://github.com/genepii/seqmet>)

Preliminary results showed that the average number of minority variants (SNPs and indels) for all samples decreased from 350 for Kit10 to 89 and 81 for Kit 12 and Kit 14 while Illumina minority variants are 67. Average number of minority indels decreased by an average factor of 10 in Kit 10 compared to Kit 12 and 14. However, it remained twice the average number of indels of Illumina. Average accuracy increased 3-fold between Kit 10 and Kit 12-14. The average recall was similar for all ONT chemistries but lower than that of Illumina.

The new ONT chemistries show an improvement in Nanopore technology usage for minority variants analysis. However, accuracy still needs to be enhanced in order to compete with Illumina technology. In future chemistry releases, ONT might be of great help to analyse minor haplotype or to unveil intra-host population dynamics.

References

1. Charre, C. and al. Evaluation of NGS-based approaches for SARS-CoV-2 whole genome characterisation. *Virus Evolution* 6, veaa075 (2020).
2. Bal, A. and al. Detection and prevalence of SARS-CoV-2 co-infections during the Omicron variant circulation in France. *Nat Commun* 13, 6316 (2022).
3. Focosi, D. and Maggi, F. Recombination in Coronaviruses, with a Focus on SARS-CoV-2. *Viruses* 14, 1239 (2022).

External quality assessment of SARS-CoV-2 variants identification and whole genome sequencing across 45 French laboratories

Arianna Tonazzoli^{*1}, Arthur Le Bars^{†1}, David Salgado^{‡1}, Naira Naouar^{§2}, Aitana Neves^{¶3}, Dillenn Terumalai^{||3}, Elaine Mc Culloch^{**4}, Alan Flisch^{††4}, Gavin Mackle^{‡‡4}, Laurence Josset⁵, Bruno Coignard⁶, Anne Bozorgan⁶, Adriana Traore⁶, Javier Castro Alvarez⁶, Bruno Lina⁵, and Jacques Van Helden⁷

¹Institut Français de Bioinformatique – Institut Français de BioinformatiqueIFB – France

²Sorbonne Université – Université Paris-Sorbonne - Paris IV – France

³Swiss Institute of Bioinformatics – Suisse

⁴Quality control for molecular diagnostics – Royaume-Uni

⁵Laboratoire de Virologie des HCL – Institut des Agents Infectieux [Lyon] – France

⁶Santé Publique France – Santé publique France – France

⁷Institut Français de Bioinformatique – Institut Français de BioinformatiqueIFB, Aix Marseille Univ – France

Résumé

In January 2021, the French government launched a national project of genomic surveillance and research about SARS-CoV-2, called EMERGEN. The consortium relies on a network of > 5000 sampling laboratories covering private medical laboratories and hospitals, and 55 sequencing laboratories equipped with Next Generation Sequencing (NGS) facilities. The sequences and the metadata are collected on a national database (EMERGEN-DB) deployed by the Institut Français de Bioinformatique (IFB), which provides access for surveillance and research. The EMERGEN project produced > 740,000 full viral genomes from January 2021 to April 2023. This rapid involvement of many laboratories and the increase of sequencing

*Intervenant

†Auteur correspondant: arthur.le-bars@france-bioinformatique.fr

‡Auteur correspondant: david.salgado@france-bioinformatique.fr

§Auteur correspondant: naira.naouar@upmc.fr

¶Auteur correspondant: aitana.neves@sib.swiss

||Auteur correspondant: dillenn.terumalai@sib.swiss

**Auteur correspondant: elainemcculloch@qcmd.org

††Auteur correspondant: alanflisch@qcmd.org

‡‡Auteur correspondant: gavinmackle@qcmd.org

Auteur correspondant: laurence.josset@chu-lyon.fr

Auteur correspondant: bruno.coignard@santepubliquefrance.fr

Auteur correspondant: anne.bozorgan@santepubliquefrance.fr

Auteur correspondant: adriana.traore@santepubliquefrance.fr

Auteur correspondant: javier.castro-alvarez@santepubliquefrance.fr

Auteur correspondant: bruno.lina@chu-lyon.fr

Auteur correspondant: jacques.van-helden@france-bioinformatique.fr

capacities prompted the consortium to assess the quality of the sequencing results and variant assignment. Therefore, a formal External Quality Assessment (EQA) was performed to quantify the performances of each sequencing laboratory, identify potential problems and suggest paths to improve the performances of consortium laboratories.

This EQA was organised into two separate phases called "challenges", consisting of thirteen inactivated SARS-CoV-2 and two negative samples in total. "Challenge 1" aimed at assessing laboratory capability in robust sample sequencing and lineage assignment for 7 samples containing one different variant each, while "Challenge 2" focused on evaluating laboratories efficiency in detecting complex viral situations, such as co-infections and recombinants.

Participating laboratories (45 French and 3 international reference labs) were assessed on the following criteria: (A) Transmission of the requested files; (B) Quality of the sequencing; (C) Consensus genome conformity; (D) Mutation call conformity and (E) Variant name conformity. Some of the quantitative scores were then used to compute three-level qualitative indicators ("Excellent", "Good" or "Needs improvement"). For most of these indicators, a large proportion of submissions obtained "Excellent" or "Good" labels. However, for the second challenge many laboratories encountered difficulties in viral lineage assignment, suggesting that these laboratories need to improve their co-infections and recombinant detection protocols.

The EQA enabled the evaluation of technical, analytical and reporting processes for the laboratories involved in French molecular surveillance of SARS-CoV-2 variants and it contributed to its enhancement. Procedure and workflows developed for this EQA will be adapted for future similar EQAs and for running a systematic quality control on all data submitted to EMERGEN-DB. This will provide regular feedback, monitor the evolution of the consortium performances and promote a continuous improvement of the quality.

Funding: this EQA was funded by ECDC/HERA/2021/007 ECD.12221; IFB is funded by the Programme d'Investissements d'Avenir (PIA), grant ANR-11-INBS-0013 and by the EMERGEN grant from Santé publique France.

Mots-Clés: external quality assessment, SARS, CoV, 2, COVID, 19, genomic surveillance

Functional annotation of dinoflagellate protein sequences and structural prediction by deep learning approach

Jérémy ROUSSEAU¹, Lucie BITTNER^{1,2} and Mathilde CARPENTIER¹

¹ Muséum National d'Histoire Naturelle – Atelier de Bio-Informatique, 45 rue Buffon, 75005, Paris, France

² Institut Universitaire de France, 1 rue Descartes, 75231, Paris, France

Corresponding Author: jeremy.rousseau2@outlook.fr

Dinoflagellates are one of the most abundant and functionally diverse groups of microbial eukaryotes. They exhibit remarkable molecular features including gigabase-sized nuclear genomes, permanently condensed chromosomes, and a 10-fold lower ratio of protein to DNA than other eukaryotic species. Genomic and transcriptomic studies from the last decade based on cultivated dinoflagellates confirmed their remarkable sequence divergence and complexity. They revealed millions of « dark » proteins, i.e., proteins that have no significant similarity to any known sequences, and that are functionally unknown [1].

In this context, we developed a new methodology allowing the functional and structural analysis of protein sequences from dinoflagellate transcriptomes and genomes. We start with a classical functional annotation with InterProScan [2]. We then build sequence similarity networks based on Diamond pairwise alignments in order to create sequence clusters that are connected components [3]. The optimal parameters of e-value, sequence alignment overlap and sequences identity were determined based on maximising the number of large components while having a high functional homogeneity score depending on Pfam annotations. Our goal is to delineate components which share sequences with similar molecular function [4]–[6]. Moreover, as structure is more conserved than sequence we predict the 3rd structure of the sequences using the last deep learning based programs, in order to also maximise structural homogeneity structure within each component and to further functionally annotate sequences [7].

Our pipeline targets the functional exploration of a genomically hyperdiverse lineage of non model eukaryotes. It aims at providing the baselines for future investigations on the metabolism, the physiology, the ecology as well as the evolution of these successful micro-organisms. Moreover, we also hope to find new protein folds that are absent from the databases.

Acknowledgements

We are grateful to the Roscoff Bioinformatics platform ABiMS (<http://abims.sb-roscoff.fr>), part of the Institut Français de Bioinformatique (ANR-11-INBS-0013) and BioGenouest network, for providing help and/or computing and/or storage resources.

The bioinformatics analyses were performed on the Core Cluster of the Institut Français de Bioinformatique (IFB) (ANR-11-INBS-0013).

References

- [1] T. G. Stephens, M. A. Ragan, D. Bhattacharya, et C. X. Chan, « Core genes in diverse dinoflagellate lineages include a wealth of conserved dark genes with unknown functions », *Sci. Rep.*, vol. 8, n° 1, Art. n° 1, nov. 2018, doi: 10.1038/s41598-018-35620-z.
- [2] P. Jones *et al.*, « InterProScan 5: genome-scale protein function classification », *Bioinforma. Oxf. Engl.*, vol. 30, n° 9, p. 1236-1240, mai 2014, doi: 10.1093/bioinformatics/btu031.
- [3] B. Buchfink, C. Xie, et D. H. Huson, « Fast and sensitive protein alignment using DIAMOND », *Nat. Methods*, vol. 12, n° 1, Art. n° 1, janv. 2015, doi: 10.1038/nmeth.3176.
- [4] E. Faure, S.-D. Ayata, et L. Bittner, « Towards omics-based predictions of planktonic functional composition from environmental data », *Nat. Commun.*, vol. 12, n° 1, Art. n° 1, juill. 2021, doi: 10.1038/s41467-021-24547-1.
- [5] H. J. Atkinson, J. H. Morris, T. E. Ferrin, et P. C. Babbitt, « Using Sequence Similarity Networks for Visualization of Relationships Across Diverse Protein Superfamilies », *PLOS ONE*, vol. 4, n° 2, p. e4345, févr. 2009, doi: 10.1371/journal.pone.0004345.
- [6] A. Meng *et al.*, « Analysis of the genomic basis of functional diversity in dinoflagellates using a transcriptome-based sequence similarity network », *Mol. Ecol.*, vol. 27, n° 10, p. 2365-2380, 2018, doi: 10.1111/mec.14579.
- [7] C. Schaefer et B. Rost, « Predict impact of single amino acid change upon protein structure », *BMC Genomics*, vol. 13, n° 4, p. S4, juin 2012, doi: 10.1186/1471-2164-13-S4-S4.

Genome analysis of SNP and SV in the admixed Creole cattle of Guadeloupe reveals new adaptative mechanisms to tropical production system

Slim Ben-Jemaa^{1,2}, Mekki Boussaha³, Philippe Bardou^{4,5}, Nathalie Mandonnet¹, and Michel Naves*¹

¹Agroécologie, génétique et systèmes d'élevage tropicaux – Institut National de Recherche pour l'Agriculture, l'Alimentation et l'Environnement – France

²Institut National de la Recherche Agronomique de Tunisie – Tunisie

³Génétique Animale et Biologie Intégrative – AgroParisTech, Université Paris-Saclay, Institut National de Recherche pour l'Agriculture, l'Alimentation et l'Environnement – France

⁴Génétique Physiologie et Systèmes d'Élevage – Ecole Nationale Vétérinaire de Toulouse, École nationale supérieure agronomique de Toulouse, Institut National de Recherche pour l'Agriculture, l'Alimentation et l'Environnement – France

⁵Système d'Information des GENomes des Animaux d'Élevage – Institut National de Recherche pour l'Agriculture, l'Alimentation et l'Environnement – France

Résumé

Multiple admixture events have shaped the genome of the Creole cattle breed from Guadeloupe (GUA) since the first time of the colonization. This breed is well adapted to the tropical environment. Through whole genome sequencing of 23 male individual representative of the population and comparisons with publicly available genomes of . 99 individuals from 25 populations across the world, we present a comprehensive characterization of the genomic variation in the GUA population. In total, 17,228,983 single nucleotide polymorphisms (SNPs), and 32821 ascertained SV defining 15258 regions, representing ~ 17% of the genome were detected. We confirm the higher level african taurine (35%) and indicine zebu (36%) ancestries, compared to the European ancestry (29 %), and we highlight the African origin of indicine ancestry, through migration. Signature of selection studied on the SNP panel, and functional analysis of the structural variants identified, highlighted various regions bearing genes with potential adaptive roles in relation to immunity, thermotolerance and physical activity, in particular with the traditional use of oxen for draught. Findings from this study provide insight into the genetic mechanisms associated with resilience traits in livestock in tropical production systems.

Mots-Clés: whole genome sequencing, cattle, adaptive traits, tropics

*Intervenant

Genome-scale essential gene discovery in a bacterium by hyper saturated transposon insertion amplicon sequencing

Domitille JARRIGE, Maria OSIPENKO and Françoise BRINGEL

Génétique Moléculaire, Génomique, Microbiologie, Université de Strasbourg UMR 7156 CNRS,
Strasbourg, France

Corresponding Authors: djarrige@unistra.fr; bringel@unistra.fr

C1 compound degradation by microorganisms is required in the carbon cycles of various ecosystems, and in case of anthropogenic C1 pollutants, a sustainable means for biological decontamination of polluted sites. Dichloromethane (DCM, CH₂Cl₂), a toxic chlorinated C1 solvent used in industry worldwide, can be degraded by the bacterium *Methylobacterium extorquens* DM4. When strain DM4 grows using DCM, chloride ions, protons and the genotoxic transient metabolite S-chloromethyl-glutathione are produced, with subsequent intracellular high chloride concentration, pH drop [1], and DNA adducts [2]. The molecular basis of the adaptive response to these multiple stresses is still poorly understood.

We compared the stress-coping strategies in this bacterium grown with DCM and with the unchlorinated C1 compound methanol (CH₃OH), using hyper saturated transposon insertion mutagenesis along strain DM4 6Mb-long genome. Around 2 million mutants grown with either methanol or DCM were obtained. Only mutants with insertions in non-essential genes were viable; these mutants had multiple insertion events along their genes. Transposon/genome junctions were amplified by semi arbitrary PCR from the mutant libraries and sequenced (Illumina 2x150bp reads). Insertion sites were mapped with TRANSIT [3], giving mean raw insertion densities of one every 13.6 nt for DCM and one every 12.1 nt for methanol. Then, with a custom Python script, insertion sites were filtered (according to the criteria presented in the poster), and insertion site distribution for each CDS was analysed. The largest uninterrupted ORF was computed for each CDS. A Z score reflecting the relative insertion density between the two growth conditions was calculated. Three categories of required genes were defined (adapted from [4]): “essential” genes had no insertions in at least 80% of their CDS; “domain-essential” genes had an uninterrupted frame between 50% and 80%; and “fitness Z” genes displayed an absolute Z score over 10. All other genes were classified as “non-essential”.

Of a total of 6,080 genes, 393 genes were required for growth with DCM, 197 with methanol and 1,159 required with both C1 compounds. User friendly-interactive figures show insertion sites along the genome, gene essentiality, CDS annotation and orientation, and recently experimentally-determined transcription start sites [5]. The “essential” analysis Python script and the data visualisation process will be published to aid future studies. We are analysing the newly-discovered required genes to confirm their essential role in the adaptive response to growth with DCM.

References

- [1] Stéphane Vuilleumier *et al.*, *Methylobacterium* genome sequences: a reference blueprint to investigate microbial metabolism of C1 compounds from natural and industrial sources, *PLoS ONE*, vol. 4, n° 5, p. e5584, 2009.
- [2] Martin F. Kayser and Stéphane Vuilleumier, Dehalogenation of dichloromethane by dichloromethane dehalogenase/glutathione S-transferase leads to formation of DNA adducts, *J Bacteriol*, vol. 183, n° 17, p. 5209-5212, 2001.
- [3] Thomas R. Ioerger, Analysis of gene essentiality from TnSeq data using Transit, in *Essential Genes and Genomes*, in *Methods in Molecular Biology*, vol. 2377. p. 391-421, NY: Springer US, 2022.
- [4] Andrea M. Ochsner, Matthias Christen, Lucas Hemmerle, Rémi Peyraud, Beat Christen, and Julia A. Vorholt, Transposon sequencing uncovers an essential regulatory function of phosphoribulokinase for methylotrophy, *Current Biology*, vol. 27, n° 17, p. 2579-2588.e6, 2017.
- [5] Bruno Maucourt, David Roche, Pauline Chaignaud, Stéphane Vuilleumier, and Françoise Bringel, Genome-Wide Transcription Start Sites Mapping in *Methylobacterium* Grown with Dichloromethane and Methanol, *Microorganisms*, vol. 10, n° 7, p. 1301, 2022.

Genome-wide CRISPR screens in B cell lymphomas reveal novel oncogenic dependencies

Camille Soun¹, Mélanie Collin¹, Sandrine Roulland¹

¹INSERM, Marseille, France

Despite its indolent nature, follicular lymphoma (FL), the second most common B-cell blood cancer in adults, remains a significant clinical burden, as the majority of patients relapse or transform into an aggressive lymphoma with poor prognosis. We currently lack an understanding of the molecular mechanisms that support abnormal FL cell proliferation/survival and of the heterogeneity of response to therapy. Among the cutting-edge technologies that have emerged in recent years, genome-wide loss-of-function CRISPR-Cas9 screens have become a powerful tool to reveal mechanisms of cell proliferation and survival. Here, we performed CRISPR screens on 10 human transformed FL cell lines using the genome-wide Brunello CRISPR library (4 sgRNA/gene, 77441 sgRNAs) and measured sgRNA abundance over 3 weeks of culture using NGS. We established an integrative analysis workflow, CRISPY, using Mageck (Model-based Analysis of Genome-wide CRISPR-Cas9 Knockout), CrisprCleanR and in-house pipelines for co-essentiality analysis. CRISPY provides a scoring metrics (Beta-score) for each gene allowing us to identify genes that are important for cell survival or proliferation and may represent potential therapeutic targets. The workflow is instrumental in determining the cellular dependencies of different tumoral cell lines in order to identify new drug targets. Comparison between genetically heterogeneous lymphoma cell lines allowed us to identify context-specific dependencies associated with genetic mutations (e.g. BCL2 apoptotic pathway or EZH2 Polycomb repressive complex, often mutated in FL), as well as recurrent dependencies across cell lines (e.g. the PI3K/SYK/AKT pathway). We also deployed a strategy to map functionally related sets of genes using co-dependency analysis of gene essentiality profiles. As expected, we uncovered known gene relationships such as the EZH2/SUZ12/EED complex and revealed novel gene relationships that we are exploring functionally. This integrative study has generated a large dataset of gene essentiality in B-cell lymphomas that we are currently validating helping us to suggest new effective targeted therapies.

Identifying genetic factors influencing the development of vascular aneurysms in Autosomal Dominant Polycystic Kidney Disease

H. LEMOINE¹, M. AUDRÉZET^{1,2}, P. HARRIS³, N. DEMOULIN⁴, R. GANSEVOORT⁵, B. KNEBELMANN⁶, Y. LE MEUR^{2,7}, THE GENKYST STUDY GROUP, THE CYSTIC STUDY GROUP, E. CORNEC-LE GALL^{1,2}

1 Univ. Brest, Inserm, EFS, UMR 1078, GGB, IBSAM, F-29200, Brest, France.

2 CHRU Brest, F-29200 Brest, France.

3 Department of Nephrology and Hypertension, Mayo Clinic, Rochester, Minnesota, USA.

4 Division of Nephrology, Cliniques universitaires Saint-Luc, Brussels, Belgium; Institut de Recherche Expérimentale et Clinique, Université catholique de Louvain, Brussels, Belgium.

5 Department of Internal Medicine, University of Groningen, University Medical Center Groningen, Groningen, the Netherlands.

6 Necker Hospital, Assistance Publique-Hôpitaux de Paris, Université Paris Cité, Paris, France.

7 Univ Brest, UMR 1227, LBAI, Labex IGO, 29200 Brest, France.

Corresponding Author: hugo.lemoine@univ-brest.fr

Autosomal dominant polycystic kidney disease (ADPKD) is the most common inherited kidney disease, with an estimated prevalence of 1/1000 to 1/2500 individuals. It is characterized by the development of multiple cysts in the renal parenchyma, leading to the progressive loss of renal function. ADPKD is mainly caused by pathogenic variants of the *PKD1* and *PKD2* genes involved in approximately 75% and 15% of cases, respectively. *PKD1* and *PKD2* encode polycystin 1 (PC1) and 2 (PC2), both membrane glycoproteins functionally expressed as a complex at the primary cilium of epithelial cells. In addition to renal phenotypes and hypertension, affected individuals may have extrarenal manifestations, including vascular phenotypes, the most common and severe of which is the development of intracranial aneurysms (IA). The prevalence of IA is five times higher in the ADPKD population than in the general population, with estimates ranging from 9 to 12%. The only risk factor for IA identified in ADPKD is a family history of IA, with a prevalence of 22% in patients with a family risk of IA, compared with 6% to 8% in the absence of a family history. The high expression of PC1 and PC2 in vascular smooth muscle tissue and endothelium and a susceptibility to develop aneurysms in *PKD1* and *PKD2* KO mouse models suggest the involvement of this dysfunctional polycystin complex in these vascular lesions. However, despite strong familial clustering of IA cases in some ADPKD pedigrees, intrafamilial variability can also be observed between side branches of some pedigrees leading to the suggestion of strong modifier genes in these vascular phenotypes. To date, no genetic factors associated with the development of IA have been identified in the polycystic population.

For this purpose, a large cohort (GENOVAS) of approximately 250 unrelated european ancestry patients with ADPKD and IA was sequenced by Whole Exome Sequencing (WES). Despite the fact that the *PKD1* gene is known to be difficult to capture, analysis of the WES data retrieved 77% of the known causal variants. The remaining causal variants were mostly found in alignment files but not called due to low read depth. More than 20 additional causal variants were also identified. Analysis of the location and type of *PKD1* and *PKD2* variants did not reveal any association with aneurysm occurrence. To identify candidate modifier genes, a gene-based burden test approach was used by comparing our large predominantly french GENOVAS cohort with ADPKD and severe IA with the french FrEx and/or european 1000Genomes control cohorts. These analyses are still in progress. Familial analysis could also be performed on several WES sequenced informative pedigrees linked to our cohort that are composed of individuals with ADPKD with or without IA.

The main objective of our study is therefore to identify the genetic factors that influence the development of IA in ADPKD patients in order to improve the pre-symptomatic screening strategy for IA and eventually allow the development of specific therapeutic strategies.

Acknowledgements

The study was supported by National Research Agency grant (ANR JCJC 2019 GENOVAS-PKD, ECLG)

MicroRNA annotation tool comparison in animal genomes

Martin Racoupeau^{1,2}, Sarah Djebali³ and Cervin Guyomar¹

¹GenPhySE, INRAE, INPT, INP-ENVT, Université de Toulouse, Castanet-Tolosan, France

²Bioinformatics Master, Université Paul Sabatier, Toulouse, France

³IRSD, Université de Toulouse, INSERM, INRAE, ENVT, Université Toulouse III - Paul Sabatier (UPS), Toulouse, France

Corresponding Authors: cervin.guyomar@inrae.fr, sarah.djebali@inserm.fr

MicroRNAs are one of the latest major breakthroughs in the realm of genetics/genomics, along with other non-coding RNAs which allowed the scientific community to further disprove the junk DNA hypothesis. They are known to repress gene expression but can also play a role in gene up-regulation by linking to promoters [1]. Beyond gene regulation, miRNAs have also been found to be able to be excreted in vesicles as well as to interact with cellular receptors [2]. With these many roles, one would expect to find refined tools able to predict and annotate miRNAs genome-wide with high confidence. This is sadly not the case, especially in non-model species such as chicken and pig. In the absence of a consensus method, we have selected three annotation tools that were not previously benchmarked together: the *mirdeep2* [3] and *miridentify* [4] heuristic tools and the graph-based and machine learning *BrumiR* suite [5]. Those tools have been run on the GENE-SWITCH (GS) chicken and pig small RNA-seq data, focusing first on the four experiments of a single tissue and developmental stage (liver, newborn). The results of each tool were post-processed to output a set of known (based on homology with *mirBase* miRNAs) and novel miRNAs, that were further compared between tools. We have found that *miridentify* was the most conservative tool (289 and 279 miRNAs predicted in chicken and pig respectively, of which 11 and 53 were novel) and *BrumiR* the least conservative (3305 and 3136 miRNAs predicted in chicken and pig respectively, of which 3043 and 2840 were novel), with *mirdeep2* lying in the middle (342 and 297 miRNAs predicted in chicken and pig respectively, of which 484 and 840 were novel). In terms of intersection, a relatively small % of known miRNAs were predicted by the three tools (16,3% and 35,4%), and no novel miRNA was predicted simultaneously by the three tools. We also found that *miridentify* and *mirdeep2* were globally more similar to each other. Predicted miRNA sizes were also much larger for *BrumiR* than for *miridentify* and *mirdeep2* (mean size of 108 bp compared to 81 and 62 for *miridentify* and *mirdeep2*). In terms of resources needed, *BrumiR* was the fastest when testing on one tissue/stage but did not scale well and was overall too resource demanding in terms of RAM (+150G) on all GS experiments. The three tools overall vary in terms of sensibility, output formats, computational resources and ease of use, and the results produced only offer a partial overlap. However, intersecting the results of multiple tools should allow to confidently annotate novel miRNAs using the full GS dataset in the future.

References

- [1] R. F. Place, L.-C. Li, D. Pookot, E. J. Noonan, et R. Dahiya, « MicroRNA-373 induces expression of genes with complementary promoter sequences », *Proc. Natl. Acad. Sci. U. S. A.*, vol. 105, n° 5, p. 1608-1613, févr. 2008, doi: 10.1073/pnas.0707594105.
- [2] M. Fabbri *et al.*, « MicroRNAs bind to Toll-like receptors to induce prometastatic inflammatory response », *Proc. Natl. Acad. Sci. U. S. A.*, vol. 109, n° 31, p. E2110-E2116, juill. 2012, doi: 10.1073/pnas.1209414109.
- [3] M. R. Friedländer, S. D. Mackowiak, N. Li, W. Chen, et N. Rajewsky, « miRDeep2 accurately identifies known and hundreds of novel microRNA genes in seven animal clades », *Nucleic Acids Res.*, vol. 40, n° 1, p. 37-52, janv. 2012, doi: 10.1093/nar/gkr688.
- [4] T. B. Hansen, M. T. Venø, J. Kjems, et C. K. Damgaard, « miRidentify: high stringency miRNA predictor identifies several novel animal miRNAs », *Nucleic Acids Res.*, vol. 42, n° 16, p. e124, sept. 2014, doi: 10.1093/nar/gku598.
- [5] C. Moraga, E. Sanchez, M. G. Ferrarini, R. A. Gutierrez, E. A. Vidal, et M.-F. Sagot, « BrumiR: A toolkit for de novo discovery of microRNAs from sRNA-seq data », *GigaScience*, vol. 11, p. giac093, oct. 2022, doi: 10.1093/gigascience/giac093.

Poster 4

Nouvel algorithme de classification agglomérative hiérarchique pour inférer un ensemble représentatif de conformations 3D d'ARN

Isaure CHAUVOT DE BEAUCHENE¹, Antoine MONIOT¹, Yann GUERMEUR¹

¹ LORIA (CNRS - INRIA - Université de Lorraine), 54000 Nancy, France

Corresponding Author: yann.guermeur@loria.fr

1 Contexte

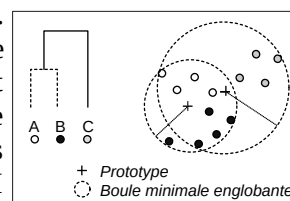
De nombreuses méthodes de modélisation moléculaire utilisent des bibliothèques 3D qui discrétisent l'espace conformationnel de molécules. Une approche classique pour les créer consiste à faire des clusters des structures expérimentales extraites de bases de données, et sélectionner un point par cluster comme *prototype*. Pour la plupart des applications, l'ensemble des prototypes doit satisfaire deux contraintes contradictoires : être représentatif de toutes les conformations existantes (pour maximiser la précision des modèles finaux) et être d'une cardinalité aussi petite que possible (pour éviter les explosions combinatoires).

2 Objectifs

A partir d'un ensemble de conformations 3D, nous voulons créer le plus petit ensemble de prototypes tel que chaque conformation soit représentée, i.e. située à moins d'un seuil de déviation d'au moins un prototype. Nous utilisons comme objet d'application un ensemble de structures de nucléotides d'ARN.

3 Méthodes

Notre problème correspond à l'inférence d'un ϵ -réseau de cardinalité minimale. Ce problème général d'optimisation combinatoire est NP-difficile et manque de méthodes dédiées. Cependant, des solutions réalisables sont dérivables, à un coût négligeable, des dendrogrammes obtenus par classification agglomérative hiérarchique (CAH). Mais les *linkage* (condition de fusion de 2 clusters) classiques sont inégalement adaptées à la tâche. Nous avons développé le algorithme CAH *radius* basé sur le calcul de *boule minimale englobante* (plus petite boule qui comprend les points d'un cluster) [1], qui nécessite un espace euclidien ou un noyau. La RMSD après superposition n'étant pas une distance (pas d'inégalité triangulaire), nous l'adaptions aux structures 3D par superpositions itératives et calculs de bornes. Nous arrêtons l'agglomération lorsque le rayon de la boule suivante est supérieur au seuil de déviation, puis nous créons le centre de chaque boule comme prototype.



4 Résultats

Nous avons d'abord effectué un test théorique sur un benchmark classique d'images de visages, et avons produit des ϵ -réseau plus petits que tous les algorithmes CAH courants. Ensuite, dans le contexte du docking par fragments d'ARN, nous l'avons appliqué aux structures 3D de trinuécléotides (~20 000 par séquence) extraites des complexes protéine-ARN de la PDB. Nous avons obtenu 566 à 1012 prototypes par séquence d'ARN avec un seuil de 1Å de RMSD, soit une réduction d'un facteur 4 à 5 par rapport au clustering conventionnel *starshape* qui conserve le point central de chaque cluster comme prototype. Cela a permis d'accélérer d'autant les calculs de docking sans diminuer la précision des poses.

5 Conclusion

Notre approche permet la création de bibliothèques 3D de cardinalité minimale pour une précision donnée. est applicable à tout ensemble de conformations 3D d'une molécule.

Acknowledgements

This work has been supported by the "RNAct" Marie Skłodowska-Curie Action (MSCA) Innovative Training Networks (ITN) H2020-MSCA-ITN-2018 (contract n° 813239).

References

- [1] A Moniot, I Chauvot de Beauchêne, Y Guermeur. Inferring Epsilon-nets of Finite Sets in a RKHS. WSOM+ - 2022, Jul 2022, Prague, Czech Republic. pp.53-62,

B-cell epitope prediction on HLA antigens using molecular dynamics simulation data

In the context of organ transplantation, unexpected production by the recipient of antibodies against donor-specific HLA antigens is the main reason for transplant loss. Thus, it is necessary to better predict B-cell epitopes* on HLA antigens in order to improve the donor-recipient matching step from a structural point of view. Although there are currently multiple tools that predict B-cell epitopes, none have shown satisfactory results on HLA antigens. This may be explained by the limited availability of HLA-antibody complex structures in the PDB [1]. Here we present a method that relies on an unprecedented dataset of short molecular dynamics (MD) simulations of 207 HLA antigens. We use hydrophobic properties, electrostatic charges, flexibility and solvent accessibility as descriptors calculated on patches sampled from MD trajectories. To overcome the lack of HLA-antibody complexes in the PDB, we leverage confirmed eplets from the HLA Eplet Registry database [2] as "ground truth" to train an Extremely Randomized Trees [3] machine learning model. This model outperforms state-of-the-art tools for B-cell epitope prediction on HLA antigens, such as DiscoTope 3.0 [4]. These results suggest the interest in using MD simulation data for the challenging task of epitope prediction.

*A B-cell epitope refers to the binding site of an antibody on the surface of the antigen.

References

[1] wwPDB consortium, « Protein Data Bank: the single global archive for 3D macromolecular structure data », *Nucleic Acids Research*, vol. 47, n° D1, p. D520-D528, janv. 2019, doi: [10.1093/nar/gky949](https://doi.org/10.1093/nar/gky949).

[2] R. J. Duquesnoy *et al.*, « Second update of the International Registry of HLA Epitopes. I. The HLA-ABC Epitope Database », *Human Immunology*, vol. 80, n° 2, p. 103-106, févr. 2019, doi: [10.1016/j.humimm.2018.11.007](https://doi.org/10.1016/j.humimm.2018.11.007).

[3] P. Geurts, D. Ernst, et L. Wehenkel, « Extremely randomized trees », *Mach Learn*, vol. 63, n° 1, p. 3-42, avr. 2006, doi: [10.1007/s10994-006-6226-1](https://doi.org/10.1007/s10994-006-6226-1).

[4] M. H. Høie *et al.*, « DiscoTope-3.0 - Improved B-cell epitope prediction using AlphaFold2 modeling and inverse folding latent representations ». *bioRxiv*, p. 2023.02.05.527174, 5 février 2023. doi: [10.1101/2023.02.05.527174](https://doi.org/10.1101/2023.02.05.527174). <https://services.healthtech.dtu.dk/services/DiscoTope-3.0/>

Comparative analysis of topologically associating domains -TAD- callers

Flavian Rique*¹, Florence Glibert¹, Cécile Dulary¹, Olivier Alibert¹§, Sophie Chantalat¹§

**first author*

§corresponding authors

¹ Functional Genomics Laboratory, National Center of Human Genomics Research, François Jacob Institute of biology, CEA

In eukaryotic cells, chromatin adopts distinct three-dimensional (3D) topologies in the nucleus, which are essential for genome functionality. In human cells, chromosomes are organized in territories and folded into compartments, inside which the chromatin is organized in large domains called topologically associating domains (TADs). TADs have been described as chromatin regions that interact more frequently within themselves than among each other. At deeper resolution, chromatin regions can form loops, which can bring into close proximity genomic sites, including regulatory regions such as promoters and enhancers. These different layers of chromatin organization are important in modulating gene expression, and in turn, altered structures can play a role in developing diseases. For example, modifications in TAD boundaries have been linked to genetic diseases and cancers, supporting their importance to maintain the genome functionality. We investigate the spatial genome organization and its impact on transcription during physiological and pathological processes by Hi-C, a high-throughput approach derived from chromosome conformation capture. Comparison of chromatin organization between different conditions (normal vs. pathological) could allow us to highlight deregulated regulatory networks governing transcriptional programs which could potentially have a critical role in the development of the disease.

Dozens of computational tools have been developed to identify TADs and their boundaries. However, a study published in 2018 showed that these tools significantly differ from each other in terms of performance and robustness. Importantly, they also display significant differences in their outputs. This lack of concordance led us to explore recent alternative methods, proposed in Cooltools and fanc pipelines, in order to see if they can outperform the previous ones. We include TopDom tool, which was the top-scoring method in the published comparison study. We found that the method proposed by Cooltools identify a larger number of TADs. Importantly, we also showed that the TAD boundaries identified by Cooltools method were more frequently enriched in either CTCF, Smc3 and Rad21, three proteins known to be required for TAD boundary formation.

Evaluation of SARS-CoV-2 Spike RBD interactions with Angiotensin Converting Enzyme 2 of multiple species using Molecular Dynamics Simulations

Damien GARCIA¹, Chloé GRIVAUD¹ and Stéphane TELETCHÉA¹

¹ Unité en Sciences Biologiques et Biotechnologies (U2SB), Nantes Université, CNRS, UMR 6286, 2 chemin de la Houssinière, 44000 Nantes, France

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

Summary:

The interaction between the SARS-Cov-2 Spike protein, in particular its Receptor Binding Domain (RBD), and the cellular receptor Angiotensin Converting Enzyme 2 (ACE2) is at the origin of the viral infection that led to the COVID-19 pandemic. Many mutations on the Spike protein appeared since its first isolation in late 2019, leading to specific Variants Of Concern (VOC) which became predominant one after the other. The objective of this study is to understand the mechanisms of interactions between the virus protein and its targeted host using Molecular Dynamics Simulations.

We have set up a robust large-scale molecular modeling methodology to study the impact of individual changes on the Spike protein leading to higher or lower recognition of a given host cell. Our methodology comprises (i) homology modeling of a Spike variant in complex with ACE2, (ii) short molecular dynamics simulations of this complex using GROMACS, (iii) MM/GBSA analysis of the interface, (iv) detailed analysis of the interface at the amino acids level.

We will present a synthetic analysis of human Variants of Concerns (VOC) and Interests (VOI) of RBD Spike proteins recognition of ACE2 receptors from multiple species (*Equus caballus*, *Felis catus*, *Mus musculus*, *Ratus norvegicus*, *Homo sapiens*). We will discuss how these analysis compare to existing literature [1, 2].

References

1. Gheeraert A, Vuillon L, Chaloin L, Moncorgé O, Very T, Perez S, Leroux V, Chauvot de Beauchêne I, Mias-Lucquin D, Devignes MD, Rivalta I, Maigret B. **Singular Interface Dynamics of the SARS-CoV-2 Delta Variant Explained with Contact Perturbation Analysis.** *J Chem Inf Model.* 2022, 62(12):3107-3122.
2. Lan J, Chen P, Liu W, Ren W, Zhang L, Ding Q, Zhang Q, Wang X, Ge J. **Structural insights into the binding of SARS-CoV-2, SARS-CoV, and hCoV-NL63 spike receptor-binding domain to horse ACE2.** *Structure.* 2022, 30(10):1432-1442.e4

Influence of IDR deamidations on Bcl-xL structure and function

Cesar Hunault, Stéphane Téletchéa

US2B, Nantes Université, CNRS, UMR 6286,
2 rue de la Houssinière, F-44000, Nantes, France

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

Platelets are small anucleated blood cells playing an essential role in blood clotting. A dramatic decrease in platelets concentration coming from accidents or diseases leads to thrombocytopenia requiring a transfusion for recovery. In human, the average lifespan of platelets is 10 days. This lifespan is under the control of 2 antagonist Bcl-2 family members: Bcl-xL for their survival and Bak for their death [1]. It is therefore essential to understand how Bcl-xL/Bak interact and how their split alter platelet life. One of our goals is to understand how these proteins interact to help extending their lifespan.

Bcl-xL undergoes two time-dependent modifications, called deamidations, that are first markers of aging but can also play a key role in the disengagement process of Bak from Bcl-xL [2,3]. These deamidations occur on asparagines 52 and 66 and can be repaired or rearranges spontaneously into aspartate and iso-aspartate. The study of Iso-Aspartate is a challenge in cellulose because these residues, called “non-natural” amino acids, cannot be encoded in proteins from DNA. In order to observe the impact of deamidations that leads to Iso-Aspartate, computational biology is a well-suited method.

We used Molecular Dynamics simulations to study these deamidation processes alone and in tandem to : (i) determine the conformation of Bcl-xL inserted into the mitochondrial outer membrane, (ii) study the consequences of the two deamidations on the conformation of the Intrinsically Disordered Region of Bcl-xL, which is known to be very flexible. We will present our recent advances on the impact of remote post-translational modifications on Bcl-xL structure, and how this can alter BAK recognition.

References

1. Mason KD, Carpinelli MR, Fletcher JI, Collinge JE, Hilton AA, Ellis S, Kelly PN, Ekert PG, Metcalf D, Roberts AW, Huang DC, Kile BT. **Programmed anuclear cell death delimits platelet life span.** *Cell* (2007) 128(6):1173-86.
2. Beaumatin F, El Dhaybi M, Lasserre JP, Salin B, Moyer MP, Verdier M, Manon S, Priault M. **N52 monodeamidated Bcl-xL shows impaired oncogenic properties in vivo and in vitro.** *Oncotarget*. (2016) 7(13):17129-43.
3. Tanriver G, Monard G, Catak S. **Impact of Deamidation on the Structure and Function of Antiapoptotic Bcl-xL.** *J Chem Inf Model.* (2022) 62(1):102-115.

Using bulk RNA-seq and iClip-seq analysis to investigate mRNA localization in synapses and uncover the role of the IMP protein

LAGHRISSE Hiba

Institute of Biology Valrose, Université Côte d'Azur

Co-authors : Hiba Laghrissi¹, Lauren Blot¹, Martina Halleger², Jernej Ule², Arnaud Hubstenberger¹, & Florence Besse¹,

¹Institute of Biology Valrose, Université Côte d'Azur

²The Francis Crick Institute, 1 Midland Road, London NW1 1AT, UK; Department of Neuromuscular Diseases, UCL Queen Square Institute of Neurology, Queen Square, London WC1N 3BG, UK.

Key words : RNA, local RNA translation, RNA transport, Bulk RNA-seq, iClip-seq, IMP protein

Abstract

Targeting of mRNA molecules to axons and dendrites plays a crucial role in neuron function, as it allows for local protein synthesis and quick changes in the synaptic proteome in response to neuronal stimulation. Our work aims to characterize the population of mRNAs specifically enriched at synapses, and at investigating the regulation and function of mRNA localization in vivo, using *Drosophila* as a model system. To this end, our team successfully isolated the synaptic content of adult *Drosophila* brain synapses, using optimized synaptosome isolation procedures. High-throughput sequencing and differential expression analysis between isolated synaptosomes and cell-body enriched fractions revealed that hundreds of mRNA species are enriched in synaptic fractions, and that these mRNAs code for various protein classes, including proteins that play a crucial role in shaping the structure and function of the synapse. To characterize the mechanisms underlying mRNA localization, we investigated the function of IMP, a conserved RNA Binding Protein (RBP) that promotes the transport of mRNAs in specific neuronal populations via its prion-like domain (PLD). We performed iCLIP-seq analysis, which allowed us to identify the RNA sequences directly bound by IMP and to compare the binding profile of Imp and IMP-dPLD. Our analysis of mRNA binding profiles as well as preferred motifs recognized by IMP will be presented. Taken together, these findings have uncovered mRNA populations enriched in synaptosomes as compared to the cell body and are starting to unravel the mechanisms involved in this process.

Enzymatic network programming in non-living biomachines for medical diagnosis

Martin DAVY¹, Patrick AMAR^{1,2} and Franck MOLINA¹

¹ Sys2Diag UMR 9005 CNRS / ALCEM, Parc Euromédecine, 1682 rue de la Valsière, CS 61003, 34184 Montpellier Cedex 4, France

² LISN, UMR CNRS 9013, Université Paris Saclay, 1 Rue Raimond Castaing, 91190 Gif-sur-Yvette, France

Corresponding author: martin.davy@sys2diag.cnrs.fr

One of the emerging field of synthetic biology is the development of non-living biomachines (NLB), artificial cells programmed to perform specific computations. NLB are composed of biological components encapsulated in a vesicle but are not alive since they have neither DNA nor replication machinery. They are frequently used in noisy biological environments but the lipid bi-layer of the membrane allow a fine control of the internal environment.

NLB have a wide range of applications; one of them is in medical diagnosis [1]. The clinical decision algorithm [2], used to perform diagnosis, is composed by multiplexed biomarker detections and revelations. NLB can be programmed to implement those detections and processing simultaneously several biomarkers to return an integrated response by producing an easily measurable and/or visible component – a readout.

It is possible to formalize a clinical decision algorithm into a logical function with Boolean functions. Biological implementation of logic gates exist in nature and can be found in transcriptional regulation network [3], as well as in metabolic and signaling networks or can be implemented with artificial enzymatic networks [4] or with artificial designed DNA [5]. Enzymatic calculation offers faster computation time than DNA and avoids the transcriptional / translational machinery needed by genetic logic gates.

The design of the artificial enzymatic networks is time consuming, considering the time to search for enzymatic reactions in databases (KEGG, BRENDA ...) and the time to build implementations. Two tools developed to facilitate these steps are presented. The first one extracts enzymatic reactions from a list of known reactions or reactions fetched from databases, choosing those that can connect biomarkers of interest to a desired readout. The second tool builds enzymatic implementations of the clinical decision algorithm by assembling the enzymatic reactions selected with the first tool.

Acknowledgements

This work was supported by grants from ANRT, CNRS and SkillCell.

References

- [1] Alexis Courbet, Eric Renard, and Franck Molina. Bringing next-generation diagnostics to the clinic through synthetic biology. *EMBO Molecular Medicine*, 8(9):987–991, 2016.
- [2] Carmi Z. Margolis. Uses of clinical algorithms. *Journal of the American Medical Association*, 249(5):627–632, 1983.
- [3] Rafael Silva-Rocha and Víctor de Lorenzo. Mining logic gates in prokaryotic transcriptional regulation networks. *FEBS Letters*, 582(8):1237–1244, 2008.
- [4] Alexis Courbet, Patrick Amar, François Fages, Eric Renard, and Franck Molina. Computer-aided biochemical programming of synthetic microreactors as diagnostic devices. *Molecular Systems Biology*, 14(4):e7845, 2018.
- [5] Lulu Qian and Erik Winfree. Scaling up digital circuit computation with DNA strand displacement cascades. *Science*, 332(6034):1196–1201, 2011.

Inferring and comparing metabolism accross heterogeneous sets of annotated genomes using AuCoMe

Arnaud BELCOUR¹, Jeanne GOT¹, Méziane AITE¹, Ludovic DELAGE², Jonas COLLEN², Clémence FRIOUX³, Catherine LEBLANC², Simon DITTAMI², Samuel BLANQUART¹, Gabriel MARKOV^{2*} and Anne SIEGEL¹

¹Univ Rennes 1, Inria, CNRS, Irisa, 35000 Rennes, France

²CNRS - Sorbonne Université - Integrative Biology of Marine Models (UMR8227) - Station Biologique de Roscoff, Place Georges Teissier, 29680, Roscoff, France

³Inria, INRAE, Université de Bordeaux, France

Corresponding Authors: arnaud.belcour@protonmail.com, anne.siegel@irisa.fr

* Presenting

Comparative analysis of Genome-Scale Metabolic Networks (GSMNs) may yield important information on the biology, evolution, and adaptation of species [1]. However, it is impeded by the high heterogeneity of the quality and completeness of structural and functional genome annotations, which may bias the results of such comparisons [2]. To address this issue, we developed AuCoMe – a pipeline to automatically reconstruct homogeneous GSMNs from a heterogeneous set of annotated genomes without discarding available manual annotations [3]. We tested AuCoMe with three datasets, one bacterial, one fungal, and one algal, and demonstrated that it successfully reduces technical biases while capturing the metabolic specificities of each organism [4]. Our results also point out shared and diverging metabolic traits among evolutionarily distant algae, underlining the potential of AuCoMe to accelerate the broad exploration of metabolic evolution across the tree of life.

Acknowledgements

We acknowledge the GenOuest bioinformatics core facility <https://www.genouest.org> for providing the computing infrastructure. We also thank Erwan Corre (ABiMS Platform) and Pauline Hamon-Giraud for fruitful discussions. This work benefited from the support of the French Government via the National Research Agency investment expenditure program IDEALG (ANR-10-BTBR-04) and from Région Bretagne via the grant SAD 2016 - METALG (9673).

References

- [1] Changdai Gu, Gi Bae Kim, Won Jun Kim, Hyun Uk Kim, Sang Yup Lee. Current status and applications of genome-scale metabolic models. *Genome Biology* 20:121, 2019.
- [2] Delphine Nègre, Méziane Aite, Arnaud Belcour, Clémence Frioux, Loraine Brillet-Guéguen, Xi Liu, Philippe Bordron, Olivier Godfroy, Agneszka P. Lipinska, Catherine Leblanc, Anne Siegel, Simon M. Dittami, Erwan Corre, Gabriel V. Markov. Genome-Scale Metabolic Networks Shed Light on the Carotenoid Biosynthesis Pathway in the Brown Algae *Saccharina japonica* and *Cladosiphon okamuranus*. *Antioxidants* 8, 564, 2019.
- [3] <https://github.com/AuReMe/aucome>
- [4] Arnaud Belcour, Jeanne Got, Méziane Aite, Ludovic Delage, Jonas Collén, Clémence Frioux, Catherine Leblanc, Simon M. Dittami, Samuel Blanquart, Gabriel V. Markov, Anne Siegel. AuCoMe: inferring and comparing metabolisms across heterogeneous sets of annotated genomes. *Genome Research*, in press, 2023. Preprint: doi:10.1101/2022.06.14.496215.

Metabolomic Modeling of Microbial Interactions for Enhanced Hydrogen Production

Xavier Marbehan¹, Eric Olmos¹, Emmanuel Guedon¹, Frantz Fournier¹, Stéphane Delaunay¹,

¹Université de Lorraine, CNRS, LRGP, F-54000 Nancy, France

Corresponding Author: xavier.marbehan@univ-lorraine.fr

Hydrogen production is a major challenge for the development of renewable energy sources. The use of microorganisms, such as *Desulfovibrio vulgaris* Hildenborough (DvH) and *Clostridium acetobutylicum* ATCC824 (Cac), can generate hydrogen without relying on fossil fuels [1]. Studies have shown that these two strains, when in physical contact, produce more hydrogen than when separated by a membrane [2]. However, the mechanisms of exchange between these organisms remain poorly understood, and their culture and analysis are complicated by their anaerobic nature and low concentrations.

To identify the metabolites responsible for the overproduction of hydrogen in this consortium, we developed a constraint-based metabolomic model (GSM) and simulated intracellular metabolic fluxes using a multi-objective modeling tool [3]. We analyzed each potentially exchanged metabolite and identified the one that leads to this overproduction of hydrogen. Our results suggest that electron exchanges between the two microorganisms are responsible for the enhanced hydrogen production, with an electron flux directed from Cac to DvH. The actors in these exchanges in our model would be the ferredoxins. This finding helps explain the observed overproduction of hydrogen and offers new perspectives for improving renewable energy production from microbial sources. Furthermore, understanding the molecular mechanisms underlying this phenomenon can pave the way for optimizing co-culture conditions and engineering microbial consortia tailored for efficient hydrogen production, ultimately contributing to a more sustainable energy landscape.

References:

- [1] Chandrasekhar, Kuppam, Yong-Jik Lee, et Dong-Woo Lee. « Biohydrogen Production: Strategies to Improve Process Efficiency through Microbial Routes ». *International Journal of Molecular Sciences* 16, no 12 (14 avril 2015): 8266-93.
- [2] Benomar, Saida, David Ranava, María Luz Cárdenas, Eric Trably, Yan Rafrafi, Adrien Ducret, Jérôme Hamelin, Elisabeth Lojou, Jean-Philippe Steyer, et Marie-Thérèse Giudici-Ortoni. « Nutritional Stress Induces Exchange of Cell Material and Energetic Coupling between Bacterial Species ». *Nature Communications* 6, no 1 (mai 2015): 6283.
- [3] X. Marbehan, F. Fournier, E. Olmos, E. Guedon et S. Delaunay (2022). Modeling of hydrogen production by a bacterial consortium, BIOPROSCALE, Berlin - Germany, 30-34 Mars 2022

Mise en place d'un outil analytique en Guadeloupe pour le dosage de la chlordécone dans le sérum humain

Mailie Saint-Hilaire*^{†1}, Didier Plumain¹, Catherine Adam², Célia Joaquim Justo², Gauthier Eppe², Stéphanie Guyomard¹, and Antoine Talarmin¹

¹Institut Pasteur de la Guadeloupe – Guadeloupe

²LEAE CART – Belgique

Résumé

La majeure partie des analyses chlordécone (CLD) sont réalisées en France Hexagonale. Une des mesures du PNACIV (mesure R5) est la mise en place d'une plateforme analytique au niveau local permettant le suivi de la CLD et de ses métabolites dans différentes matrices d'intérêts. L'objectif étant de réduire les délais et coûts d'analyse mais aussi d'accroître la capacité, la sensibilité et la précision des analyses. L'Institut Pasteur de la Guadeloupe (IPG) dispose d'un parc analytique et d'un personnel compétent pour mettre en place des outils analytiques de mesure de la CLD. Un des objectifs de l'IPG était donc le développement et la validation d'une méthode d'analyse de la CLD dans le sérum humain. Ce développement a été financé par l'ARS de Guadeloupe. La mise en place de la méthode a été réalisée en partenariat avec le LEAE-CART qui dispose d'une forte expérience sur l'analyse de la CLD dans le sérum humain. L'IPG a donc développé une méthode de type QuEChERS-LC-MS/MS pour l'analyse de la CLD dans le sérum humain. Cette nouvelle méthode rapide, sensible et peu coûteuse a été comparée à la méthode du LEAE-CART utilisée en routine. 50 échantillons de volontaires ont été analysés entre le LEAE-CART et l'IPG. A l'aide d'un test statistique d'appariement des séries, aucune différence significative n'a été observée entre les résultats des deux laboratoires. La méthode est en cours d'accréditation selon la norme NF EN 15189. La limite de quantification est de 0.06 $\mu\text{g/L}$ et cette méthode est actuellement utilisée en routine pour le dosage de la CLD dans le sang de la population guadeloupéenne. Cet outil analytique pourra aussi servir à différents projets de recherche autour de la santé de la population guadeloupéenne.

Mots-Clés: chlordécone, plateforme analytique, serum humain, Guadeloupe

*Intervenant

[†]Auteur correspondant: msaint-hilaire@pasteur-guadeloupe.fr

ODAMNet: a Python package to identify molecular relationships between chemicals and rare diseases using overlap, active module or random walk approaches

Morgane TÉRÉZOL¹, Anaïs BAUDOT^{1,2} and Ozan OZISIK¹

¹ Aix Marseille Univ, INSERM, MMG, 13385, Marseille, France

² Barcelona Supercomputing Center (BSC), 08034, Barcelona, Spain

Corresponding author: `morgane.terezol@univ-amu.fr`

In an initial study [1], we investigated the molecular relationships between vitamins A and D and Congenital Anomalies of the Kidney and Urinary Tract (CAKUT). We performed an overlap analysis between target genes of the vitamins and pathways related to CAKUT, and identified a significant overlap between vitamin A target genes and CAKUT-related pathways.

In order to systematize and extend this approach, we created ODAmNet. ODAmNet is a Python package aiming to study the molecular relationships between chemicals and rare disease pathways. We implemented three complementary approaches. The first approach is an *overlap analysis* (aka enrichment analysis). With this approach, we are looking for chemical target genes that are members of rare disease pathways. The second approach explores these molecular relationships using an *active module identification approach*. Indeed, the DOMINO tool [2] allows us to extract active modules enriched in target genes from a biological network. Then, the active modules also significantly enriched in rare disease pathways are selected. In the last approach, we measure network proximities between chemical target genes and rare disease pathways. The multiXrank Python package [3] calculates these proximities using a *random walk with restart* on a multilayer network. The multilayer network is composed of both gene-gene and gene-disease interactions. The outputs scores help us to prioritize diseases close to chemical target genes of interest.

In ODAmNet, by default the data are automatically retrieved from databases. Chemical target genes are fetched, using MeSH IDs, from the Comparative Toxicogenomics Database (CTD, <http://ctdbase.org/>). Rare disease pathways are retrieved from WikiPathways (<https://www.wikipathways.org/>). Biological networks can also be automatically downloaded from the Network Data Exchange (NDEX, <https://www.ndexbio.org/>). This allows the users to perform their analyses with up-to-date data. Importantly, the users can also provide their own target gene lists, pathways of interest or biological networks. ODAmNet is indeed open to new hypotheses and input data, not related to chemicals or rare disease. In addition, this feature makes the analyses reproducible and allows the users to give a specific version of data.

To illustrate ODAmNet, we use vitamin A as a use-case. ODAmNet retrieved from CTD and WikiPathways 2,143 vitamin A target genes and 104 rare disease pathways. We identified molecular relationships between vitamin A target genes and rare disease pathways using the overlap analysis (28 pathways) and the active module identification (19 pathways) approaches. In the random walk with restart approach, we selected the 20 best pathways based on their proximity scores. Some rare disease pathways are identified as associated with vitamin A by only one of the three approaches but, overall, more than 45% of the pathways were retrieved by multiple methods. Hence, ODAmNet exploratory analysis allows hypotheses on the molecular relationship between vitamin A and rare diseases.

Acknowledgements

We thank the members of European Joint Programme on Rare Diseases.

References

- [1] Ozan Ozisik, Friederike Ehrhart, Chris T Evelo, Alberto Mantovani, and Anaïs Baudot. Overlap of vitamin a and vitamin d target genes with cakut-related processes. *F1000Research*, 10, 2021.
- [2] Hagai Levi, Nima Rahmanian, Ran Elkon, and Ron Shamir. The domino web-server for active module identification analysis. *Bioinformatics*, 38(8):2364–2366, 2022.
- [3] Anthony Baptista, Aitor Gonzalez, and Anaïs Baudot. Universal multilayer network exploration by random walk with restart. *Communications Physics*, 5(1):170, 2022.

Study of the anaerobic degradability of chlordecone by mixed cultures

Gaëlle Gruel^{1*}, Liam Foyle^{2*}, Line Lomheim², Aiden Zhaoxiang Liu², Laurent Laquitaine¹, Suly Rambinaising¹, Naomy Duhamel¹, Ronald Ranguin¹, Corine Jean-Marius¹, Andrei Starostine², Robert Flick², Amy Li², Luz A. Puentes Jácome², Elizabeth A. Edwards², Sarra Gaspard¹

¹ Department of Chemistry, COVACHIMM2E, University of the French West Indies

² Department of Chemical Engineering and Applied Chemistry, BioZone, University of Toronto

Corresponding Author: gruel.gaelle@gmail.com

In the French West Indies, chlordecone (CLD), an active ingredient in recalcitrant organochlorine pesticides, was used extensively to control the black banana weevil from 1971 to 1993, despite its ban in the United States since 1974. Currently, significant concentrations of CLD can still be measured in some soils of Guadeloupe and Martinique, especially in banana production areas [1]. Indeed, the complex bis-homocubane structure of CLD gives it a high stability, an important hydrophobicity and a strong affinity for the organic matter of the soil, making it particularly persistent at the environmental level. This long-term pollution of environmental compartments (soils, water resources, coastal zones) impacts local food production (vegetables, livestock and seafood), thus causing human health problems and social difficulties. However, despite the sanitary and social urgency, no remediation strategy has proven satisfactory so far. The ability of certain microbial communities to degrade CLD could provide an effective and environmentally friendly bioremediation alternative. A preliminary study of the biodegradability of CLD under anaerobic conditions was carried out for two mixed bacterial cultures, one developed by the BioZone laboratory (MC1) and known for its ability to degrade organochlorine compounds [2] and the other (MC2) derived from microcosms prepared under anaerobic conditions from contaminated soils of Guadeloupe and capable of degrading lindane [3,4]. The mixed anaerobic cultures were incubated with CLD and electron donors (ethanol and acetone). Kinetics of dechlorination of CLD under anaerobic conditions and concomitant production of its metabolites were performed by LC-MS.

Reductive dechlorination of CLD under anaerobic conditions was observed in both mixed bacterial cultures. Degradation products of CLD, already identified in previous works [5-6] are detected. The major degradation product is pentachloroindene. Other metabolites, resulting from the opening of the cage structure of CLD Carboxylated -trichloroindene, and -pentachloroindene are as well detected in higher proportions in MC1, than in MC2. CLD could be quantified within both cultures. However, a research effort is still needed to obtain internal standards for the quantification of CLD metabolites.

References

1. Comte I., Pradel A., Crabit A., Mottes C., Pak L., Cattan P. Long-term pollution by chlordecone of tropical volcanic soils in the French West Indies: New insights and improvement of previous predictions, *Environmental Pollution*, 2022, 303, p. 119091.
2. Shujun Yi, Nadia Morson, Elizabeth A. Edwards, Diwen Yang, Runzeng Liu, Lingyan Zhu, and Scott A. Mabury. Anaerobic Microbial Dechlorination of 6:2 Chlorinated Polyfluorooctane Ether Sulfonate and the Underlying Mechanisms, *Environ. Sci. Technol.* 2022, 56, 907–916
3. Jacome, L.A.P., Lomheim, L., Gaspard, S., Edwards, E.A. Biodegradation of Lindane (Hexachlorocyclohexane) to Nontoxic End Products by Sequential Treatment with Three Mixed Anaerobic Microbial Cultures. *Environmental Science and Technology*, 2021, 55(5), pp. 2968–2979
4. Qiao, W., Puentes Jácome, L.A., Tang, X., Lomheim, L., Yang, M.I., Gaspard, S., Avanzi, I.R., Wu, J., Ye, S., Edwards, E.A. Microbial Communities Associated with Sustained Anaerobic Reductive Dechlorination of α -, β -, γ -, and δ -Hexachlorocyclohexane Isomers to Monochlorobenzene and Benzene (2020) *Environmental Science and Technology*.
5. L. Lomheim, L. Laquitaine, S. Rambinaising, R. Flick, A. Starostine, C. Jean-Marius, E.A. Edwards, S. Gaspard, *PLoS One*. 15 (2020) e0231219
6. M.L. Chevallier, O. Della-Negra, S. Chaussonnerie, A. Barbance, D. Muselet, F. Lagarde, E. Darii, E. Ugarte, E. Lescop, N. Fonknechten, J. Weissenbach, T. Woignier, J.F. Gallard, [S. Vuilleumier, G. Imfeld, D. Le Paslier, P.L. Saaidi, *Environ. Sci. Technol.* 53 (2019) 6133–6143

Poster 5

Exploring the epigenetic regulation of alternative splicing in the context of mouse spermatogenesis

Olivier Feudjio*¹

¹Inserm – 1Institut Cochin, INSERM U1016, UMR 8104 CNRS, Université de Paris, Commissariat à l'Énergie Atomique et aux Énergies Alternatives (CEA) - Grenoble – France

Résumé

This project investigates the relationship between alternative splicing and histone post-translational modifications during spermatogenesis. The impact of Dot1l gene knockout on alternative splicing regulation is also examined. RNA-seq data from spermatocytes and spermatid cells were analyzed using rMATS and Whippet to detect alternative splicing events. The results showed upregulation of alternative splicing in spermatocytes compared to spermatids, with intron retention showing an 81% increase in spermatocytes. The analysis of ChIP-seq data revealed the presence of H3K27ac histone marks on intron-retaining transcripts in spermatocytes.

Mots-Clés: RNA, seq, Alternative splicing, histone PTMs, H3K79 methylation, rMATS, Whippet, ADLIN workspace

*Intervenant

Result reproducibility and practice standardisation is a common issue in the field of bioacoustics, as well as science in general. Even between members of the same team, programming languages or conventions can differ, making the reuse of a code tedious and time-consuming, even for its original author. The collaborative project OSmOSE¹ (Open Science meets Ocean Sound Explorers), started in 2017 at ENSTA Bretagne (Brest, France), aims to make progress on this issue in the field of underwater passive acoustic monitoring. On the programming side, the OSmOSE team has developed several specialized pipelines to structure data, generate spectrograms and automatically detect biological signals of interest. Our code development process tries to stick at best with F.A.I.R. principles. This includes the distribution of our codes as an open source python package, the writing of a complete and detailed documentation and a test suite covering most of the code. Our tools have also been made user-friendly with prefilled jupyter notebooks for common cases, while being highly configurable for advanced users. It was also made suitable to be run on the cluster Datarmor of IFREMER using the pbs job management system. This current presentation provides details on the development of this suite of tools, and illustrates benefits on different scientific use-cases in underwater passive acoustic monitoring.

¹ <https://osmose.ifremer.fr/>

Impact du formaldéhyde sur les données WGS PCR Free issues d'échantillons tumoraux FFPE dans le cadre du plan France Médecine Génomique 2025

Marine Rouillon¹, Mélanie Letexier^{1,3}, Alice Moussy³, Jasmin Cevost¹, Edouard Turlotte¹, Alain Viari^{1,2}, Jean-François Deleuze^{1,3}

¹ Centre de Référence, d'Innovation, d'Expertise et de transfert (CRefIX), US 039 CEA/INRIA/INSERM, 91000, Evry, France

² Synergie Lyon Cancer, centre Léon-Bérard, 69008, Lyon, France

³ Centre National de Recherche en Génomique Humaine (CNRGH), Institut de Biologie François Jacob, Direction de la Recherche Fondamentale, CEA, 91000, Evry, France

Corresponding Author: marine.rouillon@cnrgh.fr

1. Contexte et Objectifs

La production de génomes entiers dans le cadre du plan FMG2025 répond à deux objectifs : augmenter le taux diagnostic pour les patients et fournir des données de qualité pour la recherche. Pour satisfaire ce double-enjeu, la fiabilité des données produites est essentielle. Ainsi, pour limiter les biais techniques, le protocole initial de séquençage choisi pour les plateformes fut le WGS PCR Free à partir d'échantillons tumoraux cryopréservés (FF). Cependant, la majorité des échantillons tumoraux disponibles sont fixés et inclus en paraffine (FFPE). Les modifications chimiques apportées à l'ADN par la fixation peuvent entraîner une fragmentation aléatoire et une modification de séquence qui peuvent fausser la détection des variants et le nombre de copies du génome. C'est pourquoi le CRefIX, centre R&D du plan, a évalué les répercussions de la fixation des tissus tumoraux FFPE, comparé aux tissus FF, en séquençage WGS PCR-Free. L'objectif est de déterminer si ces modifications moléculaires de l'ADN ont un impact sur l'interprétation diagnostique finale.

2. Matériels et Méthodes

Quatre blocs miroirs (FF vs FFPE), et les PBMC associés, de patients atteints de cancer du poumon ont été séquençés avec le kit TruSeq DNA PCR-free d'Illumina sur NovaSeq 6000. Dans un souci d'uniformité, les données ont été normalisées à 80X de profondeur de séquençage. La détection des variants somatiques est effectuée par Mutect2 (GATK v4.1.7.0), l'analyse des CNVs par FACETS (bioconductor/3.13), la détection des SVs par Manta (v1.0.3) et les statistiques/figures sont générées à l'aide du logiciel R (v4.1.1).

3. Résultats et Discussion

Le rendement des bibliothèques FFPE est largement inférieur à celles des bibliothèques FF mais leur quantité d'ADN reste suffisante pour le séquençage et leur qualité est équivalente. Seule l'homogénéité de couverture apparaît comme déviante lors du contrôle qualité.

Le processus de préparation des tissus FFPE génère des artéfacts de séquence qui sont interprétés à tort comme des mutations. Ces faux positifs sont présents lors de l'appel de variants et lors de l'analyse de variabilité du nombre de copies (CNAs). Ces dommages semblent avoir une distribution aléatoire et uniforme sur le génome. Des mutations C>T sont retrouvées de façon prédominantes dans les SNVs spécifiques aux échantillons FFPE mais l'interprétation clinique finale du profil mutationnel n'est quant à lui pas impacté.

Après analyse, les artéfacts de SNVs peuvent être corrigés en ayant une profondeur de lecture adéquate et en cohérence avec le pourcentage de cellularité. La nécessité de travailler à partir d'un tissu suffisamment riche en cellule tumorale (>=40%) et une profondeur minimale de 80X est requise. Un filtre sur la fréquence des variants (VAF) est aussi efficace et augmente le F1-score. Contrairement aux analyses de SNVs, la fiabilité de détection des gains et des pertes est remise en cause, le signal est perturbé par les grandes fluctuations du log2R dans le FFPE.

4. Conclusion

La fixation des tissus par le FFPE ne perturbe pas de façon majeure l'analyse biologique finale des variants et des SVs lorsque des filtres adéquats sont appliqués. Cependant, l'analyse des CNAs présente des discordances visibles, des investigations supplémentaires seront menées afin d'éclaircir ce phénomène.

Le CRefIX bénéficie d'une aide du PIA-ANR dans le cadre du plan France 2030 (ANR-18-INBS-0001).

ISO-seq transcriptomics analysis for polyploid genomes

Mathis POCHON¹, Martine DA ROCHA, Sophie MANTELIN¹,
Cyril VAN GHELDER¹, Céline LOPEZ-ROQUES², Marine SALLABERRY²,
Céline VANDECASTEELE², Etienne GJ DANCHIN¹, KARINE ROBBE-SERMESANT¹

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

² GeT-PlaGe, 24 chemin de borde rouge Auzeville, 31326, Castanet-Tolosan, France

Corresponding Author: karine.robbe-sermesant@inrae.fr

Transcriptomics has been extensively studied with short reads RNA-seq for the last decade. However, it can be limiting when studying complex transcriptomes such as those of polyploid organisms. High accuracy single molecule long-read isoform sequencing (Iso-Seq) opens new perspectives towards improving our knowledge of transcriptomes and alternative splicing [1]. The technology is based on PacBio sequencing and subreads alignment consensus for high accuracy. First promising studies were made for diploid or haploid organisms [2], but difficulties for polyploid organisms still need to be overcome. Indeed, auto-polyploids, resulting from whole genome duplication or allo-polyploids resulting from inter-species hybridization yield a set of genes in multiple copies that can be highly similar and particularly difficult to discriminate.

Here, we propose a workflow for Iso-seq analysis adapted for polyploid organisms. As polyploid species model, we are using a plant parasitic nematode, *Meloidogyne incognita*, having a triploid genome as a result of two hybridization steps with two of the three copies being very similar [3]. Typical tools and pipelines tend to merge homeolog transcripts especially during a polishing step. As a consequence, high identity thresholds are necessary during all isoform alignment steps and polishing needs to be limited. This polyploid adapted iso-seq workflow (using isoseq3, minimap2, TAMA) [4,5,6] will improve the quality of genome annotation at the isoform level and will facilitate polyploid differential expression analysis.

References

1. Rory Stark. RNA sequencing: The teenage years. *Nature Reviews Genetics* 20, n°11:631-56, 2019.
2. Ali Ali. PacBio Iso-Seq improves the rainbow trout genome annotation and identifies alternative splicing associated with economically important phenotypes. *Frontiers in Genetics* 12: 683408, 2021.
3. Romain Blanc-Mathieu. Hybridization and polyploidy enable genomic plasticity without sex in the most devastating plant-parasitic nematodes. *PLoS Genetics* 13 (6): e1006777, 2017.
4. Sébastien Guizard. Nf-Core/Isoseq: Simple gene and isoform annotation with PacBio Iso-Seq long-read sequencing. *Bioinformatics (Oxford, England)*, btad150, 2023.
5. Heng Li. Minimap2: Pairwise alignment for nucleotide sequences. *Bioinformatics (Oxford, England)* 34, n° 18: 3094-3100, 2018.
6. Richard I Kuo. Illuminating the dark side of the human transcriptome with long read transcript sequencing. *BMC Genomics* 21, n° 1 : 751, 2020.

QuaDS: Qualitative-Quantitative Descriptive Statistics

Andréa Bouanich^{*†1}, Angelina El Ghaziri², Pierre Santagostini², Alix Pernet³, Claudine Landès¹, and Julie Bourbeillon^{‡2}

¹IRHS - Équipe BIDEfl (Bioinformatics for Plant Defense Investigations) – Institut de Recherche en Horticulture et Semences – France

²IRHS - Équipe ImHorPhen (Imagerie pour l’Horticulture et le Phénotypage) – Institut de Recherche en Horticulture et Semences – France

³IRHS - Équipe GDO (Génétique et Diversité des plantes Ornementales) – Institut de Recherche en Horticulture et Semences – France

Résumé

QuaDS is a bioinformatic pipeline written in Python 3. It is the Python 3 version to the *catdes* R function from the *FactoMineR* package with some extras. The goal of this pipeline is to obtain descriptive statistics and an interactive visualisation of different clusters of individuals.

Mots-Clés: Clustering, R function to a Python one, descriptive statistics, interactive visualisations

*Intervenant

†Auteur correspondant: andrea.bouanich@inrae.fr

‡Auteur correspondant: julie.bourbeillon@agrocampus-ouest.fr

sRNA-pipe: a Nextflow-based pipeline for small RNA analysis

MATTHIAS ZYTNICKI¹ and Hoang-Giang PHAM¹

¹ Applied Mathematics and Computer Science Unit of Toulouse (MIAT)-INRAE, 24 chemin de Borde Rouge, 31320, Toulouse, France

Corresponding Author: hoang-giang.pham@inrae.fr

Motivation:

Small RNA refers to a class of RNA molecules that are typically less than 200 nucleotides in length. There are several types of small RNA, some of the most well-documented types of small RNA includes: microRNAs (miRNAs), small interfering RNAs (siRNAs), and piwi-interacting RNAs (piRNAs). They play a key role in regulating gene expression, have been implicated in a variety of diseases.

Today, next-generation sequencing (NGS) has revolutionized our understanding of the field of small RNA analysis, providing a powerful tool for identifying and quantifying small RNAs in a high-throughput and cost-effective manner.

However, small RNA analysis faces some challenges. For example, in mapping and quantification step, small RNA sequences are often short and can map to multiple locations in the genome. This can make it difficult to accurately align sequencing reads to a reference genome or to estimate the level of expression of a gene given the number of reads that map to this gene. Accurately identifying differentially expressed small RNAs can also be challenging, particularly when dealing with lowly expressed or lowly abundant small RNAs.

To the best of our knowledge, open-source pipelines use tools that don't deal well with the multi-mapping reads. The objective of this work is to develop a pipeline which could solve that problem in small RNA analysis.

Workflow:

The workflow management software used is Nextflow [1] which is a powerful and flexible platform for building and running reproducible and scalable computational pipelines for data analysis. The pipeline can be used to analyze small RNA sequencing data obtained from organisms with reference genome. It takes a sample sheet and FASTQ files as input, and perform quality control (QC), trimming, mapping, annotation, and differential gene expression. It uses the following tools to map, quantify multi-mapping reads, and detect differential expression of small RNAs such as srnaMapper [2], mmannot [3], mmquant [4], and srnadiff [5].

Perspectives:

The pipeline will be sharing publicly at <https://github.com/phamhoanggiang262/srna-pipe>. The current pipeline is being benchmarked on plant data (*Arabidopsis thaliana*) (1 000 000 first reads of each sample). Then, the pipeline will be tested on animal and human data. After that, because the pipeline is being created under the nf-core guidelines, we hope that we will contribute to some modules of the nf-core (the community of NextFlow) repository. Alternatively, the pipeline could be merged with the pipeline nf-core/smrnaseq.

- [1] P. Di Tommaso, M. Chatzou, E. W. Floden, P. P. Barja, E. Palumbo, and C. Notredame, "Nextflow enables reproducible computational workflows," *Nat. Biotechnol.*, vol. 35, no. 4, pp. 316–319, Apr. 2017, doi: 10.1038/nbt.3820.
- [2] M. Zytnecki and C. Gaspin, "srnaMapper: an optimal mapping tool for sRNA-Seq reads," *BMC Bioinformatics*, vol. 23, no. 1, p. 495, Nov. 2022, doi: 10.1186/s12859-022-05048-4.
- [3] M. Zytnecki and C. Gaspin, "mmannot: How to improve small-RNA annotation?," *PloS One*, vol. 15, no. 5, p. e0231738, 2020, doi: 10.1371/journal.pone.0231738.
- [4] M. Zytnecki, "mmquant: how to count multi-mapping reads?," *BMC Bioinformatics*, vol. 18, no. 1, p. 411, Sep. 2017, doi: 10.1186/s12859-017-1816-4.
- [5] M. Zytnecki and I. González, "Finding differentially expressed sRNA-Seq regions with srnadiff," *PloS One*, vol. 16, no. 8, p. e0256196, 2021, doi: 10.1371/journal.pone.0256196.

KiNext: A pipeline for the identification and classification of protein kinases

Elisabeth HELLEC¹, Charlotte CORPOREAU¹, Flavia NUNES² and Alexandre CORMIER³

¹ Ifremer-PHYTNESS-UMR LEMAR, Centre Bretagne - route de Sainte-Anne, 29280 Plouzané, FRANCE

² Ifremer-LEBCO, Centre Bretagne - route de Sainte-Anne, 29280 Plouzané, FRANCE

³ Ifremer-SeBimer, Centre Bretagne - route de Sainte-Anne, 29280 Plouzané, FRANCE

Corresponding Author: elisabeth.hellec@ifremer.fr

Protein kinases are a family of proteins typically involved in signal transduction, allowing organisms to detect and respond to biotic or abiotic environmental change [1]. Kinases contain conserved domains, thought to be homologous from bacteria to humans. Given their important role in organismal physiology, including acclimation to environmental conditions, identifying the kinome of any organism could be useful. An increasing availability of genomes makes it possible to examine and compare the complement of protein kinases across organisms throughout the tree of life.

In this context, the objective of the work is to develop a pipeline respecting the FAIR principles (findable, accessible, interoperable and reusable) in order to search, classify and determine the phylogeny of the protein kinases of any organism from its proteome (predicted from genome annotation). The first step of this pipeline is to identify protein kinases among the proteome and separate the "conventional" eukaryotic protein kinases (ePKs) from "atypical" protein kinases (aPKs) using Hidden Markov Models (HMMs), generated from the catalytic domains of known kinases from *H. sapiens*, *D. melanogaster* and *C. elegans* (Kinbase). The second step further classified ePKs into eight families based on sequence similarity of catalytic domains. In order to achieve this classification, a library containing one HMM for each kinase sub-family [2] was used against the ePKs identified in the previous step. Once these ePKs have been classified a phylogenetic analysis is carried out.

KiNext was first validated on the predicted proteome of *Crassostrea gigas* [3] and *Ostreococcus tauri* [4], organisms for which a kinome had previously been compiled. The pipeline recovered all kinases previously identified for this species, and included 10 additional kinases. Then, it was also used on the newly acquired genome of the honeycomb worm *Sabellaria alveolata* [5][6]. Identifying and classifying the kinomes of marine organisms from different habitats will allow comparison of how these families are related (in number, isoforms and expression through transcriptomes) to the lifestyle of the species, whether it is subtidal, intertidal, temperate or tropical. By increasing the number of marine species for which the genome is available, this work will allow us to compare a large number of kinomes between species and to have a better understanding of the nature of the biological responses that would be impacted by climate change such as global warming or ocean acidification.

References

1. Manning G, D. B. Whyte, R. Martinez, T. Hunter, and S. Sudarsanam. "The Protein Kinase Complement of the Human Genome." *Science* 298, no. 5600: e0155435, December 6, 2002.
2. Miranda-Saavedra, D, and Barton, G.J. Classification and functional annotation of eukaryotic protein kinases. *Proteins* 68, 893-914. [PubMed](#), 2007.
3. Epelboin Yanouk, Laure Quintric, Eric Guévelou, Pierre Boudry, Vianney Pichereau, and Charlotte Corporeau. "The Kinome of Pacific Oyster *Crassostrea Gigas*, Its Expression during Development and in Response to Environmental Factors." *PLOS ONE* 11, no. 5: e0155435, mai 2016.
4. Hindle, M. M. et al. The reduced kinome of *Ostreococcus tauri*: core eukaryotic signalling components in a tractable model species. *BMC Genomics* 15, 640 (2014).
5. Hellec E. Identification et étude comparative du kinome chez l'huître creuse *Crassostrea gigas* et le ver marin *Sabellaria alveolata*. Master 1 thesis, Université de Rennes 1, 2022.
6. Robert J. Assemblage du génome de l'annélide polychète *Sabellaria alveolata*. Master 2 thesis, Université Rennes 1, 2020.

Jobim 2023 Poster

AskoR, an R package for easy RNA-Seq data analysis

Susete ALVES CARVALHO*, Kévin GAZENGEL*, Sylvain MASANELLI, Anthony BRETAUDEAU, Stéphanie ROBIN, Stéphanie DAVAL and Fabrice LEGEAI
INRAE - IGEPP, Domaine de la motte, 35650, Le Rheu, France
*contributed equally

Corresponding author: kevin.gazengel@inrae.fr

To make the process of transcriptomics data easier, and to guarantee the reproducibility of the analyses, we have developed AskoR, which is an R tool to achieve a suite of statistical analyses and graphical outputs from gene expression data obtained by high-throughput sequencing (RNA-seq). From raw counts generated by mapping and counting tools, it allows to filter and normalize data, to check the consistency of samples, to perform differential expression tests, to execute GO-term and KEGG enrichments, and to define co-expression clusters corresponding to expression patterns between experimental conditions. The edgeR package [1] was chosen for differential expression analyses, topGO [2] for GO-term enrichments and coseq [3,4] for co-expression clusters identification. DiffGraph [5] and igraph [6] packages are actually tested to improve AskoR tool with gene network analyses. Users can define a large number of parameters (about 60) as, for example, significance thresholds for statistical tests, algorithms for each tool or the generation of specific graphs. The tool has the advantage of being flexible : on the one hand, novices users can apply the default parameters defined on the basis of those commonly used, and on the other hand, experienced users can go further in the analyses by adapting these settings. All analysis steps automatically and quickly generate a large number of tables and figures in an output folder as summary tables for each step, expression heatmaps, volcano-plots, Venn diagrams or Upset graphs (less than 25 minutes for 16 transcriptomes and 20,000 genes). AskoR can therefore be used to analyze gene expression from RNA-seq experiments but can also be extrapolated to SmallRNA-seq or metagenomics data, as well as any other experiment that leads to the generation of a count table (excepted enrichment analyses). Finally, the tool produces outputs compatible with the AskOmics tool [7] developed to integrate complex data.

AskoR can be downloaded from Askomics GitHub (<https://github.com/askomics/askoR>) and used in your R environment or is directly accessible via the national scale Galaxy server hosted by the IFB (<https://usegalaxy.fr/>). It is planned to make it available on CRAN.

References

1. Robinson MD. *et al.* edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics*, 26(1):139-140. 2010.
<https://doi.org/10.1093/bioinformatics/btp616>
2. Alexa A. and Rahnenfuhrer J. topGO: Enrichment Analysis for Gene Ontology. R package version 2.52.0. 2023.
3. Rau A. and Maugis-Rabusseau C. Transformation and model choice for co-expression analysis of RNA-seq data. *Briefings in Bioinformatics*, 19(3)-425-436. 2018.
4. Godichon-Baggioni A. *et al.* Clustering transformed compositional data using K-means, with applications in gene expression and bicycle sharing system data. *Journal of Applied Statistics*. 2018.
<https://doi.org/10.1080/02664763.2018.1454894>
5. Xiao-Fei Zhang *et al.* DiffGraph: an R package for identifying gene network rewiring using differential graphical models. *Bioinformatics*, Volume 34, Issue 9, Pages 1571–1573. 2018.
<https://doi.org/10.1093/bioinformatics/btx836>
6. Csardi G. and Nepusz T. The igraph software package for complex network research. *InterJournal, Complex Systems*, 1695. 2006.
7. Garnier X. *et al.* AskOmics: a user-friendly interface to Semantic Web technologies for integrating local datasets with reference resources. *JOBIM, Nantes, France*. pp.1. hal-02401750. 2019.

Method optimization for Rapid Pathogen Identification in Lower Respiratory Infection in Low Resource Settings.

Jean¹ ARNAUD, Jocenais² PIERRE, Herald² JEAN, Timothy¹ FORD

¹ Ford Lab, 3 Solomont Way #005-#007, Lowell, MA, 01854, USA

² Clinique UNDH, 69, Rue Paul Eugène Magloire, Hinche - Haiti, P.O Box: 1594, Haiti

Corresponding Author:

Jean_Arnaud@uml.edu, Jean.chem.Arnaud@gmail.com, jarnaud2009@gmail.com

ABSTRACT

Lower Respiratory Infections (LRI) are a major cause of mortality and morbidity in Haiti, particularly among children and older adults. Additionally, the rise of Antimicrobial Resistance (AMR) has made the management of LRIs more challenging and has increased the need for rapid, accurate, and affordable diagnostic methods. Results from traditional microbiology can take several days, and many pathogens will not grow on laboratory media. The results are also seldom useful for clinical diagnosis of microbial resistance. The need for accurate and early pathogen identification extends beyond LRI. It covers cases of emergency sepsis, site infection in cases of surgeries, monitoring of intervention against diseases such as diphtheria or malaria and more. High throughput sequencing offers a solution by using DNA as a fingerprint to identify pathogens and polymicrobial infections. Identification of genes for antimicrobial resistance can support doctors in prescribing the effective medication and also limiting the risk of increasing antimicrobial resistance.

Once out of reach, advances in sequencing technology have significantly reduced associated costs and have rendered improved diagnostics in developing countries a possibility. The MinION is a portable sequencer that has been used in many different environments such as real-time surveillance of the Ebola virus, sequencing of DNA in space, monitoring of *Salmonella* in food, and many other applications. This project is informed by existing and publically available protocols. The project will aim to improve extraction capacity through low-cost polymeric ionic liquid coupled with zirconia bead beating to eliminate the need for costly and at times toxic extraction kits. It will use digestive enzymes and PCR to amplify signals. Additionally, it will seek to improve the capacity for offsite analysis through the integration of a portable, low-cost raspberry pi supercomputer for data analysis.

In conclusion, the project will demonstrate the potential of a sensitive, robust and low-cost protocol in improving patient outcome through rapid diagnosis for LRI.

Acknowledgements

Special acknowledgements to the entire staff of Clinique UNDH for their continued engagement and insight to this project.

References

1. Charalampous, T., Kay, G.L., Richardson, H. *et al.* Nanopore metagenomics enables rapid clinical diagnosis of bacterial lower respiratory infection. *Nat Biotechnol* **37**, 783–792 (2019). <https://doi.org/10.1038/s41587-019-0156-5>
2. Nacham, O., Clark, K.D., Anderson, J. L., Extraction and Purification of DNA from Complex Biological Sample Matrices Using Solid-Phase Microextraction Coupled with Real-Time PCR. *Analytical Chemistry* **99**(15), 7813-7820 (2016). <https://doi.org/10.1021/acs.analchem.6b01861>
3. Heidorn, Ben. Build Your Own Super Computer with Raspberry Pis. Udemy. www.udemy.com/course/build-your-own-super-computer-with-raspberry-pi/

Executive Summary

Background and Context: Lower Respiratory Infections (LRI) and Antimicrobial Resistance (AMR) pose significant challenges in Haiti, particularly for vulnerable populations. Current diagnostic methods are time-consuming and often fail to identify resistant strains. Advances in sequencing technology offer a solution by using DNA as a fingerprint for pathogen identification. This project aims to develop an accurate, affordable, and accessible diagnostic method for LRI and AMR in Haiti.

Objectives:

1. Develop a diagnostic method for LRI that detects both the infection and resistance mechanisms in a single test.
2. Improve patient management by providing timely and accurate results.
3. Increase availability and accessibility of the diagnostic method.
4. Disseminate findings and impact on healthcare.

Method Development:

1. Collect sputum samples from patients with suspected LRI.
2. Deplete host DNA and purify the sample matrix.
3. Enrich microbial DNA using bead beating.
4. Extract DNA using cost-effective and rapid protocols and compare to standard method.
5. Perform PCR and sequencing using ONT's MinION and Flongle.
6. Analyze data using real-time software and optimize computational performance.
7. Identify resistance mechanisms through targeted gene analysis.

Validation:

1. Validate the diagnostic method using archived samples and mock bacterial samples.
2. Evaluate sensitivity, specificity, and predictive values.
3. Develop a validation plan, train lab personnel, and ensure ethical compliance.
4. Establish data management and sharing protocols.
5. Review with local authorities and seek regulatory approval.

Scaling up and Sustainability:

1. Develop a plan for long-term sustainability.
2. Explore financing options and partnerships for widespread availability.
3. Monitor and evaluate the diagnostic method's performance and maintenance.

Integrating the miRnome into multiple omics of Richter Transformation: insights into the development of aggressive lymphomas

Romain PIUCCO, Ghislain FIÉVET, Pierre FEUGIER, Julien BROSEUS, and Sébastien HERGALANT

Nutrition, Genetics and Environmental Risk Exposure (NGERE) – Inserm U1256, Campus Biologie Santé, Nancy, France

Corresponding author: romain.piucco@univ-lorraine.fr

Chronic Lymphocytic Leukemia (CLL) is the abnormal proliferation of small, mature B cells in lymphoid tissues, blood, and the bone marrow. CLL is a slow-progressive often asymptomatic disease manageable in most cases. However, it may transform into an aggressive lymphoma with a Diffuse Large B-Cell Lymphoma (DLBCL) histology in 90% cases, called Richter Transformation (RT). RT is resistant to chemotherapies and associates with a dismal prognosis: the median overall survival is < 12 months [1]. Most RT, but not all, derive from one of the preceding CLL clones (80% cases). Although multiple studies explored the genetic characteristics and the mechanisms involved in the transformation, none of the corresponding CLL histological and molecular features are currently usable to anticipate it. Moreover, discriminating between RT and *de novo* DLBCL is impossible without knowledge of the preceding CLL. This information is often missing, and to date, only a few classifiers of DNA methylation and gene expression data are available to infer the CLL origin of RT [2].

MiRNA (microRNA) are small epigenetic modifiers interfering with the translation process of many transcripts, including those involved in DNA accessibility and chromatin remodelling [3]. In CLL, deregulation of the miRnome, the omic layer quantifying every miRNA, and an overall disrupted epigenetic landscape have been observed, but few data are available at RT stage. The goals of this project are thus to explore RT miRnomes for further insights into the epigenetic mechanisms underlying the disease, improve existing classifiers discriminating RT from DLBCL, and predict outcome in CLL and in previously unclassified lymphomas. Here we will analyse more than 50 RT, CLL prone to RT and DLBCL miRnomes by small RNA-sequencing, perform differential and discriminating analyses, and integrate the data into a multi-omics setup covering other molecular layers available for these pathologies: coding and non-coding transcriptomes, overlapping proteomes and phosphoproteomes, DNA methylomes, exomes and copy-number variations [2,4,5]. To this aim, we are designing a specific miRNA-sequencing integrative pipeline encompassing a mixture of unsupervised methods (mixOmics) [6], deep-learning tools (custOmics) [7], and in-house approaches for paired and independent samples.

References

1. Rossi D, Spina V, Deambrogi C, et al. The genetics of Richter syndrome reveals disease heterogeneity and predicts survival after transformation. *Blood*. 2011;117(12):3391-3401.
2. Broséus J, Hergalant S, Vogt J, et al. Molecular characterization of Richter syndrome identifies *de novo* diffuse large B-cell lymphomas with poor prognosis. *Nat Commun*. 2023;14(1):309. 2023.
3. Treiber T, Treiber N, Meister G, Regulation of MicroRNA Biogenesis and Its Crosstalk with Other Cellular Pathways. *Nat Rev Mol Cell Biol* 2019, 20 (1), 5–20.
4. Ferreira PG, Jares P, Rico D, et al. Transcriptome characterization by RNA sequencing identifies a major molecular and clinical subdivision in chronic lymphocytic leukemia. *Genome Res*. 2014;24(2):212-226.
5. Kulis M, Merkel A, Heath S, et al. Whole-genome fingerprint of the DNA methylome during human B cell differentiation. *Nat Genet*. 2015;47(7):746-756.
6. Rohart F, Gautier B, Singh A, Lê Cao KA. mixOmics: An R package for 'omics feature selection and multiple data integration. *PLoS Comput Biol*. 2017;13(11):e1005752.
7. Benkirane H, Pradat Y, Michiels S, Cournède PH. CustOmics: A versatile deep-learning based strategy for multi-omics integration. *PLoS Comput Biol*. 2023;19(3):e1010921

Poster 6

The Carbohydrate-Active enZyme database: literature, functions, subfamilies and recent Python scripts to update it

Matthieu Boulinguez^{*1} and Elodie Drula^{*†1}

¹Architecture et fonction des macromolécules biologiques – Aix Marseille Université, Centre National de la Recherche Scientifique, Institut National de Recherche pour l’Agriculture, l’Alimentation et l’Environnement – France

Résumé

Thirty years have elapsed since the emergence of the classification of carbohydrate-active enzymes in sequence-based families that became the CAZy database and website (www.cazy.org) over 20 years ago. In the era of large scale sequencing and high-throughput Biology, it was important to examine the position of this specialist database that is still deeply-rooted in human curation. The three primary tasks of the CAZy curators are (i) to semi-manually annotate the modularity of daily released sequences (Genbank daily-nc); (ii) to create novel families following literature discoveries and (iii) to capture any additional functional characterization informing on the diversity of specificity in each family. In our recent publication (Drula et al., 2022, NAR Database Issue), we summarized the increase in novel families and annotations during the last eight years. We also presented major changes, developed to facilitate taxonomic navigation (Krona charts) and the download of the entirety of CAZy annotations. We also highlighted the considerable amount of work that accompanies the capture of biochemical data from the literature. The survival of CAZy relied on numerous collaborations with biological experts in the last decade, exploiting comparative genomics analysis with CAZyme expertise. We notably invested important efforts in the curation of fungal genomes from the 1KFG program (JGI), as well as in exploiting bacterial operons termed Polysaccharide Utilization Loci (PUL). Some of these collaborations will be highlighted, such as (i) the isofunctional but structurally/evolutionary distinct SusG proteins in human gut *Bacteroides* (with the Koropatkin lab, University of Michigan, USA), or (ii) the SYNBIOGAS project on landfill microbiomes (J. McDonalds, Bangor University, Wales) in which new isolates were compared with reference genomes. The exploration of always more genomes, among the diverse (quality of) genomes at NCBI, can be challenging. A python script was implemented to rank candidate genomes by scaffolds, taxonomic diversity, and putative CAZymes, prior to curation/integration in CAZy. Another script follows NCBI taxonomy regular renaming of species, 12 per week during the past year, which can sometimes lead to the loss of information (notably strain details). Finally, reported functions in CAZy follow the Enzyme Commission (EC) system, e.g. 3.2.1.21 for β -glucosidase, while we frequently had to create a temporary/incomplete number to avoid awaiting for the EC, such as 2.4.99.- for heptosyltransferase. A python script was written to watch the EC releases; and detect CAZy temporary/incomplete numbers that would correspond.

*Intervenant

†Auteur correspondant: elodie.drula@univ-amu.fr

Mots-Clés: glycogénomique, base de données CAZy, domaines protéiques, opérons bactériens, annotation fonctionnelle de microbiomes

BIPAA, Bioinformatics Platform for the Agroecosystems Arthropods.

Stéphanie Robin^{*†1,2}, Anthony Bretaudeau^{1,2}, and Fabrice Legeai^{1,2}

¹Institut de Génétique, Environnement et Protection des Plantes – Université de Rennes 1, Agrocampus Ouest, Institut National de Recherche pour l’Agriculture, l’Alimentation et l’Environnement – France

²Institut de Recherche en Informatique et Systèmes Aléatoires – Université de Rennes 1, Institut National des Sciences Appliquées - Rennes, Université de Bretagne Sud, École normale supérieure - Rennes, Institut National de Recherche en Informatique et en Automatique, CentraleSupélec, Centre National de la Recherche Scientifique : UMR6074, IMT Atlantique Bretagne-Pays de la Loire – France

Résumé

BIPAA (BioInformatics Platform for the Agroecosystems Arthropods) (<https://bipaa.genouest.org>) is a bioinformatics platform from the French National Institute for Agriculture, Food and Environment (INRAE). It is located in Rennes (France) and is integrated in the GenOuest infrastructure (<https://www.genouest.org>). It is dedicated to assist genomics and post-genomics programs developed on insects associated to agroecosystems. More than seven hundred users are currently listed on BIPAA.

#Data analysis. BIPAA is supporting a network of scientists from various french labs for analyzing their genomics data. Depending on the needs, it implies the personal guidance for developping scripts, running complex workflows on a computing cluster or on Galaxy servers. We collaborate with many biology labs for computing and analyzing heterogeneous data covering many bioinformatics topics such as genome assembly and annotation, expression analysis, non-coding RNA characterization, genomes comparison, variant identification, or genomics and epigenomics data integration.

#Databases. BIPAA is the home of several public reference databases hosting multiple insect genomes: AphidBase (for aphids), LepidoDB (for lepidopterans) and ParWaspDB (for parasitoid wasps). More than 40 genomes are currently available online. For each genome, a collection of web applications allows to explore reference genome or transcriptome assemblies and annotations (e.g. genome browser, gene reports), to analyze this data (e.g. dedicated Galaxy server, specific web applications), and to collect new scientific knowledge (e.g. manual curation of annotations using Apollo).

#Collaborations. Often in collaboration with the GenScale and Dyliss teams in INRIA/Irisa in Rennes (France), BIPAA is engaged in various research programs involving bioinformatics skills. For example, an user-friendly web application for integrating and querying heterogeneous data (AskOmics) or a tool for long non-coding RNA annotation (FEELnc) were developed in this context. BIPAA is also associated with 3 national networks of INRAE : BAPOA (Biologie Adaptative des Pucerons et Organismes Associés), ADALEP (Adaptation

*Intervenant

†Auteur correspondant: stephanie.robin@inrae.fr

à l'environnement biotique chez les lépidoptères) and REacTION (Réseau d'échange sur les mécanismes Épigenétiques qui façonnent les interacTIONs) networks. It is also involved in international consortia or various insect genome sequencing projects like i5k, an initiative to sequence the genomes of 5000 arthropods or ERGA (European Reference Genome Atlas) project.

#Trainings. Moreover, frequent training sessions in partnership are organized with the GenOuest bioinformatics platform or the BARIC (Bioinformatique pour l'Analyse, la Représentation et Intégration de Connaissances) community of INRAE.

Mots-Clés: Arthropods, genomics, data analysis, databases, Galaxy, trainings

MoPSeq-DB: the reference database and visualisation platform for marine mollusc pathogens genomes

Clémentine Battistel^{*1}, Jean-Christophe Mouren¹, Benjamin Morga¹, Camille Pelletier¹, Lydie Canier¹, Céline Garcia¹, Isabelle Arzul¹, Laura Leroi², Patrick Durand², Germain Chevignon¹, and Maude Jacquot^{†1}

¹RBE-ASIM – Institut Français de Recherche pour l’Exploitation de la Mer (IFREMER) – France

²IRSI-SeBIMER – Institut Français de Recherche pour l’Exploitation de la Mer (IFREMER) – France

Résumé

Pathogen surveillance and diagnostic methods are constantly evolving. Sequencing can provide critical information to diagnose diseases and inform control and mitigation strategies by identifying genetically distinct pathogen variants that may have different host reservoir species or geographic distributions.

During the last decade, with the development of high throughput sequencing methods, reference laboratories and research groups studying marine mollusc diseases have advanced considerably in sequencing-based analyses. However, data management is still unconventional as the community lacks dedicated databases and tools, despite a considerable increase in data volume.

We therefore developed a user-friendly web-platform, called MoPSeq-DB, which references curated genomic data related to mollusc pathogens. It gives users opportunities to navigate through data, interactively visualise genomic structure and variation, provide integrated analysis tools, and allow to download data in various file formats. Since marine bivalve molluscs can be affected by viral, bacterial and eukaryotic pathogens, MoPSeq-DB is designed to be used with a large panel of genome particularities (e.g. size, architecture).

MoPSeq-DB, is an open-source tool based on the Python web-framework Django enabling convenient and fast sequencing data exploration and visualisation in an intuitive and user-friendly way, particularly for non-bioinformaticians. It has minimal hardware requirements and is easy to install, host, and update.

While MoPSeq-DB folder structure enforces systematic yet flexible storage of genomic data of bivalve mollusc pathogens, including associated metadata, the platform could easily be declined to pathogens of any other organisms. The application can be deployed using a Docker container, and runs on all modern browser engines (Firefox, Chrome, Safari).

Source code and documentation are available at <https://gitlab.ifremer.fr/bioinfo/mopseq-db>.

A public web server will be online at: <https://mopseq-db.ifremer.fr> by mid 2023.

^{*}Intervenant

[†]Auteur correspondant: maude.jacquot@ifremer.fr

Mots-Clés: marine mollusc pathogens, database, visualisation platform, genome, sequencing based analyses

InDeepNet a web application to assist drug design

Fabien Mareuil^{*†1}, Vincent Mallet , Alexandra Moine-Franel , Luis Checa Ruano ,
Guillaume Bouvier , and Olivier Sperandio[‡]

¹Hub Bioinformatique et Biostatistique – Institut Pasteur [Paris] – C3BI 25-28, rue du Docteur Roux,
75724 Paris cedex 15, France

Résumé

Protein-protein interactions (PPIs (1)) are crucial components of many biological pathways and are increasingly being targeted in drug discovery projects, particularly those aimed at treating infectious diseases. However, developing drugs that aim to interact with PPIs is a complex task that demands extensive effort to identify appropriate candidates and assess their potential as therapeutic targets. This involves determining the role of PPIs in disease pathways, characterizing them experimentally, and predicting their ability to interact with other proteins or be modulated by drugs.

InDeep (2) is a tool that utilizes deep learning to predict functional binding sites within proteins that could serve as binding sites for protein epitopes or future drugs. By leveraging a curated dataset of PPIs, InDeep can make enhanced predictions of functional binding sites on experimental structures and within molecular dynamics trajectories. Our benchmarking results demonstrate that InDeep outperforms existing ligandable binding site predictors when assessing PPI targets, as well as conventional targets. This tool therefore provides new opportunities to assist drug design projects by identifying relevant binding pockets at or near PPI interfaces.

To make InDeep more accessible, we have developed InDeepNet, a web application that allows users to easily use InDeep on their own protein structures, molecular dynamics trajectories, or on protein structures from public databases such as PDB or the AlphaFold2 repository. InDeepNet manage asynchronous InDeep calculation on a GPU Kubernetes cluster, and offers advanced visualization and interaction capabilities for proteins and predicted pockets by an original integration of Mol* (3) within a React interface. Additionally, InDeepNet offers a complete REST API implemented with the Django REST framework.

References

Rachel Torchet, Karen Druart, Luis Checa Ruano, Alexandra Moine-Franel, H el ene Borges, Olivia Doppelt-Azeroual, Bryan Brancotte, Fabien Mareuil, Michael Nilges, Herv e M enager, Olivier Sperandio, The iPPI-DB initiative: a community-centered database of protein–protein interaction modulators, *Bioinformatics*, Volume 37, Issue 1, 1 January 2021, Pages 89–96, <https://doi.org/10.1093/bioinformatics/btaa1091>

^{*}Intervenant

[†]Auteur correspondant: fmareuil@pasteur.fr

[‡]Auteur correspondant: olivier.sperandio@pasteur.fr

Vincent Mallet, Luis Checa Ruano, Alexandra Moine Franel, Michael Nilges, Karen Druart, Guillaume Bouvier, Olivier Sperandio, InDeep: 3D fully convolutional neural networks to assist in silico drug design on protein–protein interactions, *Bioinformatics*, Volume 38, Issue 5, March 2022, Pages 1261–1268, <https://doi.org/10.1093/bioinformatics/btab849>

David Sehnal, Sebastian Bittrich, Mandar Deshpande, Radka Svobodová, Karel Berka, Václav Bazgier, Sameer Velankar, Stephen K Burley, Jaroslav Koča, Alexander S Rose, Mol* Viewer: modern web app for 3D visualization and analysis of large biomolecular structures, *Nucleic Acids Research*, Volume 49, Issue W1, 2 July 2021, Pages W431–W437, <https://doi.org/10.1093/nar/gkab314>

Mots-Clés: deeplearning, structural bioinformatic, web application

ABiMS: Analysis and Bioinformatics for Marine Science

Lorraine BRILLET-GUÉGUEN^{1,2}, Gildas LE CORGUILLÉ¹, Mark HOEBEKE¹, the ABiMS team, and Erwan CORRE¹

¹ CNRS - Sorbonne Université - Plateforme ABiMS - Station Biologique de Roscoff, Place Georges Teissier, 29680, Roscoff, France

² Sorbonne Université, CNRS, Integrative Biology of Marine Models (LBI2M), Station Biologique de Roscoff (SBR), 29680, Roscoff, France

Corresponding Author: erwan.corre@sb-roscoff.fr

Des domaines tels que la biologie des systèmes, la modélisation des réseaux ou l'analyse des données NGS constituent un véritable défi en termes de calcul scientifique. Dans un contexte de production de données de biologie marine à haut débit et de traçabilité des analyses, le développement d'une infrastructure de calcul scientifique est une étape essentielle pour la production de connaissances.

La plateforme ABiMS (Analysis and Bioinformatics for Marine Science) de la Station biologique de Roscoff répond aux besoins des chercheurs et des chercheuses en biologie marine et, plus largement, des sciences de la vie. Créée en 2008, elle est l'une des plateformes nationales de l'Institut français de bioinformatique (IFB). Elle est également un Centre de Données et de Services *in situ* du pôle ODATIS, de l'infrastructure de recherche Data Terra. ABiMS fait partie du réseau IBISA et est membre du GIS BioGenouest.

La plateforme met au service de la communauté une infrastructure de calcul et de stockage (2500 CPU , 2.5 Po), ainsi qu'une palette de compétences, un catalogue de services organisés autour de 5 activités :

- Ingénierie logicielle : développement d'interfaces web couplées à des bases de données, centrées aussi bien sur des données de type séquence que sur des données d'observation
- Gestion de données : FAIRisation ou accompagnement à la FAIRisation, mise en accès de jeux de données FAIRisées, publication de jeux de données DOIisés dans des entrepôts thématiques
- Analyse bioinformatique : analyse de données (Assemblage et annotation de genome, transcriptomes, metagenomes etc, analyse de diversité), etc.
- E-infrastructure : cluster de calcul, interfaces Galaxy, JBrowse, Apollo, R, espace de stockage ou outils bioinformatiques. Environnement pour l'analyse, l'annotation et l'hébergement de données de génomiques
- Support : demande de support pour l'utilisation des ressources de la plateforme (logiciels, données, etc.)
- Formation : formation aux méthodes et logiciels bioinformatiques

Pour assurer la qualité de nos services, nous avons mis en place un système de gestion de la qualité basé sur la norme ISO 9001, qui a été initialement certifié en 2014 et qui est approuvé à la norme ISO 9001:2015 depuis décembre 2017.

Grâce à ses nombreuses interactions avec les unités de recherche et en tant que membre du Réseau national des ressources informatiques de l'IFB et en tant que centre de données et de service , ABiMS est impliquée dans de nombreux projets de recherche (une vingtaine par an), dont les impacts nationaux et européens concernent les activités de bioanalyse, de développement logiciel et d'e-infrastructures.

Joint transcriptome and translome analysis: a reproducible pipeline

Julie RIPOLL¹, Fati CHEN², Céline MANDIER¹ and Eric RIVALS^{1,3}

¹ MAB - LIRMM - Université Montpellier - CNRS, Montpellier, France
² ADVANSE - LIRMM - Université Montpellier - CNRS, Montpellier, France
³ IFB-CORE - Institut Français de Bioinformatique, Montpellier, France

Corresponding Author: julie.ripoll@lirmm.fr

RNA sequencing (RNA-seq) is often used to elucidate the regulation of gene expression, as it provides an in-depth view of transcribed RNAs and a relative quantification of each gene or transcript. However, several studies have found that variations in RNA transcript levels do not necessarily correlate well with protein levels [1], suggesting that translation also plays a key role. To study the role of translation in regulating gene expression, the translome can be studied by selecting ribosome-bound RNAs, releasing ribosomes, and then sequencing the selected population of RNAs as in RNA-seq. This approach, called polysome sequencing (POL-seq), is a census assay that provides relative expression levels of genes/transcripts during translation. Using the workflow manager Snakemake [2] and Conda [3] environments, we developed a bioinformatics pipeline to jointly analyze these transcriptome and translome fractions.

To promote reusability of individual steps, we propose a lightweight wrapper system to facilitate interoperability on different hardware without requiring an Internet connection. Moreover, it preserves the command inspection in Snakemake's dry-run mode.

This pipeline is divided into two parts for primary and secondary analysis. The primary analysis part consists of 4 steps: i) quality control, ii) cleaning of the sequencing reads, iii) mapping of the reads to the reference genome, and iv) counting of the mapped reads per gene. The secondary analysis part encapsulates all statistical analyses to estimate the differential expression of genes and then perform functional enrichment analyses (*i.e.* gene ontology and pathway databases).

For the statistical part, all samples are normalized together to avoid any comparison problems between fractions. We then combine the comparisons of each fraction, visualized by a scatter plot, and filter the data according to their transcriptional and translational expression levels. This allows us to classify differentially expressed mRNAs into eight categories. These categories represent mRNAs that are regulated by transcription only, by translation only, or by both transcription and translation. In the case of conjoint regulations, we determine whether these regulations have combined or opposing effects.

We will illustrate this pipeline with results from a publication with our collaborators [4].

Acknowledgements

This project was supported by the *Ligue Contre le Cancer* and by the Institut National du Cancer (INCa, FluoRib grant – PLBIO18-131). The development of pipeline for the community is supported by ANR-11-INBS-0013236.

References

- [1] Alexis Battle, *et al.* Impact of regulatory variation from rna to protein. *Science*, 347(6222):664–667, Feb 2015.
- [2] Johannes Köster, Sven Rahmann. Snakemake--a scalable bioinformatics workflow engine. *Bioinformatics*. 2012 Oct 1;28(19):2520-2. doi: 10.1093/bioinformatics/bts480. Epub 2012 Aug 20. Erratum in: *Bioinformatics*. 2018 Oct 15;34(20):3600. PMID: 22908215.
- [3] *Anaconda Software Distribution*. (2020). *Anaconda Documentation*. Anaconda Inc. Retrieved from <https://docs.anaconda.com/>
- [4] Gabriel Therizols, *et al.* Alteration of ribosome function upon 5-fluorouracil treatment favors cancer cell drug-tolerance. *Nature Communications*, 13(1):173, January 2022.

MOAL: Improving the Reproducibility of OMIC Bioanalysis

Florent Dumont*^{†1} and Luciana Oliveira*^{‡2}

¹Université Paris-Saclay - fac de pharmacie – Université Paris-Sud - Université Paris-Saclay – France

²ADLIN-Science – ADLIN-Science – France

Résumé

Abstract

Exploiting omic data has become a usual task within many medical and biological research laboratories. Indeed the increasing number of technological platforms and their attached services, the lessening of costs, and the publication of raw data impulse the creation of new standards in analysis methodology. In this context, numerous efficient bioinformatics tools are available to make preprocessing and generate raw data matrix. Typically, the following step is the normalization that are, in a similar way, assists by numerous mathematical methodologies available in programming or graphical ways. At this step, we usually must deals with a numerical normalized matrix unballasted for technical device noise and ready to make downstream analysis. Here, we define as " bioanalysis " finals analysis steps leading to biological interpretation where results must be scaffolding by infographics integrated to scientific literature data.

*Currently, bioanalysis suffers do not have a simple, standardized and automatized workflows to apply on a laptop for reproducible research at laboratory. In this context, we developed **moal** (Multi Omic Analysis at Lab) (1) a R package including an easy-to-use omic function in order to response to this lack.*

Briefly, the workflow needs normalized data, statistical experimental design and symbol annotations. First, we apply quality control and unsupervised analysis on global data for all input factors. Then we apply analysis of variance model and compute post hoc Tukey pairwise comparisons tests with biological sense and their combination pattern with fold-change. After multiple test correction, the workflow filters the results using a threshold gradient on fold-changes to make downstream analysis. All data subsets are used to generate venn and cluster analysis. All created lists of significant features from all subset analysis are then used for global functional analysis: MSigDB geneset (2) database is used for enrichment analysis (canonical pathways, disease patterns, upstream regulators...), topGO (3) tools is used to analyse ontological terms and interaction network are created using STRING DB(4).

*To maintain a collaborative and transparent spirit in exploiting the results, we have created a fruitful private-public partnership with **ADLIN Science**(5). This digital health-tech company is developing an innovative solution for data management and multi-omics analysis.*

*Intervenant

[†]Auteur correspondant: florent.dumont@universite-paris-saclay.fr

[‡]Auteur correspondant: Luciana@adlin-science.com

ADLIN platform promotes the open science initiative, following fair principles to guarantee the security and traceability of scientific research. ADLIN workspace is a user-friendly, integrated environment that assists and guides the user in building and managing projects with different levels of complexity. Individual modules facilitate a wide range of tasks, such as data structuration, bioinformatics analysis, etc., carried out in parallel.

References

- (1) doi zenodo (in progress)
- (2) Liberzon A, Birger C, Thorvaldsdóttir H, Ghandi M, Mesirov JP, Tamayo P. The Molecular Signatures Database (MSigDB) hallmark gene set collection. *Cell Syst.* 2015 Dec 23;1(6):417-425.
- (3) Alexa A, Rahnenfuhrer J (2022) topGO: Enrichment Analysis for Gene Ontology.. R package version 2.50.0.
- (4) Szklarczyk et al. *Nucleic Acids Res.* 2015 43(Database issue):D447-52
- (5) ADLIN Science. <https://adlin-science.com/>. Accessed: 2022-05-12.

Mots-Clés: omic omics bioanalysis workflow R packages reproducibility functional analysis

Le CRefIX, le centre de Référence, d'Innovation, d'eXpertise et de transfert du plan France Médecine Génomique 2025

Violette Turon^{1,3}, Kévin Gorrichon¹, Margaux Gras¹, Marine Rouillon¹, Mélanie Letexier^{1,3}, Edouard Turlotte¹, Jasmin Cévest¹, Christine Michel^{1,3}, Alain Viari^{1,2}, Jean-François Deleuze^{1,3}

¹ Centre de Référence, d'Innovation, d'eXpertise et de transfert (CRefIX), US 039
CEA/INRIA/INSERM, 91000, Evry, France

² Synergie Lyon Cancer, centre Léon-Bérard, 69008, Lyon, France

³ Centre National de Recherche en Génomique Humaine (CNRGH), Institut de Biologie
François Jacob, Direction de la Recherche Fondamentale, CEA, 91000, Evry, France

Corresponding Author: melanie.letexier@cnrgh.fr

Le CRefIX (Centre de Référence, d'Innovation, d'eXpertise et de transfert) est l'une des trois structures clés du plan France Médecine Génomique 2025 avec les plateformes de séquençage et le collecteur analyseur de données, visant à déployer l'analyse génomique dans l'offre de soins. Il a pour missions :

- 1) d'établir des référentiels et standards biologiques et bio-informatiques, nécessaires à assurer la reproductibilité et l'interopérabilité des données ;
- 2) de stimuler l'innovation et d'accélérer le transfert technologique en lien avec le tissu industriel ;
- 3) de participer au développement d'une filière industrielle nationale en médecine génomique.

Créé début 2019 comme une Unité Mixte de Service (US39) associant le CEA, l'INSERM et l'INRIA, le CRefIX est hébergé au Centre National de Recherche en Génomique Humaine (CNRGH) au sein de la Genopole d'Évry. Il est doté d'une plateforme test reflétant l'ensemble des équipements utilisés dans les plateformes de production de séquençage diagnostique du plan (Novaseq 6000, ...), permettant ainsi un transfert optimum, du CRefIX vers les plateformes, des technologies développées ou évaluées.

Le CRefIX répond à des questions scientifiques et techniques du plan en émettant des recommandations, propose des projets innovants pour anticiper les besoins de la médecine génomique et assure une veille technologique pour évaluer toute amélioration disponible :

- Il a initié des projets de R&D, notamment avec le plan génomique anglais, Genomics England, sur l'apport du séquençage longue lecture pour l'augmentation du taux diagnostique via l'analyse des variants structuraux impliqués, par exemple, dans certains cancers ou maladies rares comme la déficience intellectuelle.
- Le développement de nouveaux matériels de référence biologiques en oncologie est également un projet de cette collaboration internationale.
- L'évaluation d'un nouveau kit WGS PCR free, réduisant la quantité d'ADN de départ, avant le déploiement sur les plateformes (intérêt majeur pour les microbiopsies en oncopédiatrie) a été réalisée.
- Actuellement, seules les biopsies tumorales cryopréservées sont incluses dans le plan. L'introduction de tissus tumoraux fixés dans du formol (FFPE) est en cours de mise en place sur les plateformes. Le formol a un impact sur l'ADN/ARN et donc sur la fiabilité des résultats de séquençage. Le CRefIX a mené des travaux sur l'impact des protocoles d'extraction et de préparation de bibliothèques sur les résultats de séquençage ainsi que sur l'apport de traitements bio-informatiques sur ces données contenant des artefacts causés par le formol.
- Concernant les projets collaboratifs public-privé, le CRefIX s'est rapproché d'un consortium regroupant entreprises de biotechnologie et d'informatique pour étudier la mise en place d'un workflow pour le suivi de patients cancéreux à partir de l'analyse de l'ADN tumoral circulant provenant de biopsies liquides.
- Egalement, un projet d'évaluation de séquenceurs de la technologie MGI a démarré fin 2021.

Ainsi, le poster se focalisera sur des réalisations concrètes menées par le CRefIX dans le cadre de ses missions de R&D et d'innovation pour le plan FMG2025.

Le CRefIX bénéficie d'une aide financière du PIA-ANR dans le cadre du plan France 2030 pour réaliser ses missions (ANR-18-INBS-0001).

ETBII: a new IFB school on Integrative Bioinformatics

Lucie Khamvongsa-Charbonnier¹, Anaïs Baudot^{2,3,4}, Samuel Chaffron⁵, Olivier Dameron⁶, Alban Gagnard⁷, Carl Herrmann⁸, Delphine Potier⁹, Morgane Terezol², Jimmy Vandel¹⁰, Olivier Sand^{1,10} and Hélène Chiapello^{1,11}

¹ CNRS, Institut Français de Bioinformatique, IFB-core, UMS 3601, Évry, France

² Aix Marseille Univ, INSERM, MMG, UMR1251, Marseille, France

³ CNRS, Marseille, France

⁴ Barcelona Supercomputing Center, Barcelona, Spain

⁵ Nantes Université, École Centrale Nantes, CNRS, LS2N, UMR 6004, F-44000 Nantes, France

⁶ Université de Rennes, Inria, CNRS, IRISA - UMR 6074, F-35000 Rennes, France

⁷ Nantes Université, CNRS, INSERM, Institut du Thorax, 8 quai Moncousu, Nantes F-44000, France

⁸ Institute for Pharmacy and Molecular Biotechnology & BioQuant, University Heidelberg, 69120 Heidelberg, Germany

⁹ Centre de Recherche en Cancérologie de Marseille (CRCM); CNRS, Aix Marseille Univ, INSERM, Institut Paoli-Calmettes, Marseille, France

¹⁰ Plateforme Bilille, Université de Lille, CNRS, Inserm, CHU Lille, Institut Pasteur de Lille UAR 2014 - US41 - PLBS, F-59000 Lille, France

¹¹ Université Paris-Saclay, INRAE, MaIAGE, 78350 Jouy-en-Josas, France

Corresponding Author: lucie.khamvongsa-charbonnier@france-bioinformatique.fr

The French Institute of Bioinformatics (IFB) is the National Bioinformatics Infrastructure that provides support, deploys services, organises training and carries out innovative developments for the life science communities. According to our inquiries targeting life scientists and bioinformaticians, almost all teams and units express the need for training. The most requested skills are related to NGS analysis, biostatistics and data analysis (including machine learning and AI), and bioinformatics skills, especially related to the recent field of integrative bioinformatics.

In this poster we will present a new IFB training school named ETBII that has been designed in 2022 to enhance integrative bioinformatics skills at the national level and fill gaps in the existing french training activity landscape. The first edition took place in January 2023 in Fréjus (France) with a double objective: (i) the enhancement of bioinformaticians theoretical and practical skills, (ii) the development of shared training material on this topic. The target audience of this first edition was mainly bioinformaticians from IFB platforms and teams wishing to improve their knowledge and skills on this subject and to contribute to the constitution of pedagogical material in order to deliver future training sessions on this theme ("train the trainers" approach). The poster will present the content and organisation of the ETBII first training session, with a specific focus on difficulties related to this topic and innovative pedagogical aspects deployed during this first session. We will also present future projects and actions related to the development of training resources in integrative bioinformatics, both in France and in other Elixir platforms.

The Institut Français de Bioinformatique (IFB) is funded by the Programme d'Investissements d'Avenir (PIA), grant Agence Nationale de la Recherche number ANR-11-INBS-0013

Heterogeneous data integration; Knowledge representation; Semantic web; Functioning of complex biological systems; Multiscale analysis and modelling; Multivariate analysis; Dimension reduction; Data integration; Cluster; Cloud;

DeepOmics Submission, a plug-in tool to facilitate the submission of meta-omics data to the ENA

Baptiste ROUSSEAU¹, Véronique JAMILLOUX¹, Thomas DENECKER³, Hélène CHIAPELLO^{2,3} and Ariane BIZE¹

¹ Université Paris-Saclay, INRAE, PROSE, 92761 Antony, France

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

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

Corresponding author: `ariane.bize@inrae.fr`

INRAE PROSE has coordinated the development of DeepOmics [1], an information system dedicated to meta-omics datasets in the field of environmental biotechnology. DeepOmics offers the possibility to upload, query and export 16S metabarcoding data from several environmental biotechnologies, together with many relevant associated metadata, especially regarding operating conditions and process design. To date, DeepOmics has all the data and metadata necessary for submission of the datasets to international repositories such as the European Nucleotide Archive (ENA) [2], but submission remains a manual and rather complex process.

In collaboration with the French Institute of Bioinformatics (IFB), we designed and developed a R-shiny [3] plug-in application, **DeepOmics Submission**, to prepare and automate raw sequencing data submission to the ENA. We will also guarantee the quality of the associated metadata by an IFB data brokering application prototype. This tool will promote the sharing and publication of environmental meta-omics datasets in agreement with open science and FAIR data principles.

Acknowledgements

We acknowledge N. Raidelet, A. Gramusset (DSI-INRAE) and G. Heinrich (ID2L) for helpful discussions and feedback.

References

- [1] Ariane Bize, Guillaume Perreal, Aurélie Gramusset, Marion Predhumeau, Cédric Midoux, Nicolas Raidelet, Valentin Loux, Yannick Fayolle, Patrick Dabert, and Théodore Bouchez. Deepomics, a digital environmental engineering platform for meta-omics data. In *Posters proceedings of the 2020 edition of JOBIM*, page 25. JOBIM, 2020.
- [2] Rasko Leinonen, Ruth Akhtar, Ewan Birney, Lawrence Bower, Ana Cerdeno-Tárraga, Ying Cheng, Iain Cleland, Nadeem Faruque, Neil Goodgame, Richard Gibson, Gemma Hoad, Mikyung Jang, Nima Pakseresht, Sheila Plaister, Rajesh Radhakrishnan, Kethi Reddy, Siamak Sobhany, Petra Ten Hoopen, Robert Vaughan, Vadim Zalunin, and Guy Cochrane. The European Nucleotide Archive. *Nucleic Acids Research*, 39(suppl_1):D28–D31, 10 2010.
- [3] Winston Chang, Joe Cheng, JJ Allaire, Carson Sievert, Barret Schloerke, Yihui Xie, Jeff Allen, Jonathan McPherson, Alan Dipert, and Barbara Borges. *shiny: Web Application Framework for R*, 2023. R package version 1.7.4.9002.

athENA : a Nextflow pipeline for sequencing data brokering to ENA

Pauline Auffret¹, Laura Leroi¹, Julie Clément², Alizée Bardon¹, Alexandre Cormier¹, Cyril Noël¹, Bernard de Massy³, Patrick Durand¹

1.Ifremer, IRSI, SeBiMER Service de Bioinformatique de l'Ifremer, F-29280 Plouzané, France

2.IHPE, Univ Montpellier, CNRS, IFREMER, Univ Perpignan Via Domitia, Perpignan, France

3.Institut de Génétique Humaine (IGH), CNRS, Univ Montpellier, Montpellier, France

Corresponding Author: pauline.auffret@ifremer.fr

To follow the FAIR principles (Findable, Accessible, Interoperable, Reusable) [1], research data must open up to the world, which is a prerequisite for publishing articles in peer-reviewed journals. Depending on their nature and source, several open public repositories can host and publish research data. In the specific case of sequencing data, the International Nucleotide Sequence Database Collaboration (INSDC) coordinates and synchronizes three dedicated repositories DDBJ-DRA (Japan), EMBL-EBI-ENA (UK), and NCBI-SRA (USA) that offer free and open access to raw reads, alignments, assemblies and functional annotation.

Raw reads submission to ENA can be done using three different procedures: interactive filling out of web forms and data upload; use of Webin-CLI command-line program [2] or entire programmatic submission using XML files and cURL (or equivalent) protocol. For researchers, following these procedures appears time-consuming, unclear for the uninitiated, and difficult to handle. Moreover, although minimal metadata standards are required upon submission, they are far from sufficient to FAIRly describe a sequencing dataset. To help researchers with data submission, i.e. data brokering (in others words the act of submitting data on behalf of another person), few initiatives have been developed so far. The easy-to-use interfaced Galaxy ENA Upload tool [3] and the suite of Python scripts *ena-submit* tool [4] facilitate the submission to ENA from an Excel metadata template, but lack of portability and metadata validation before submission.

Here we present athENA, a data brokering command-line pipeline inspired by *ena-submit* tool, which provides i) an extended Excel metadata template composed of 6 thematic sheets in compliance with ENA metadata model, using drop-downs lists to help filling out fields based on controlled vocabulary, ii) checking that enough metadata is filled in according to selected thematic sample checklist, numeric/date fields format control ; iii) file conversion from Excel to one XML file per ENA metadata objects : study, sample, experiment, run ; iv) secure parallel upload of raw sequencing reads to ENA FTP server and v) programmatic submission of XML objects to ENA test or production server with 2-year maximum embargo. The 3 main steps of the pipeline execution, 1) XML generation and validation, 2) data upload, 3) object submission, can be run independently if needed. For step 3, user selects the best-suited submission action: VALIDATE (checks object conformity without actually submitting them), ADD (adds a new object) or MODIFY (modifies a pre-existing object).

athENA is a free and open source pipeline [5], optimized to run on High Performance Computing clusters. It requires Nextflow DSL2 and one of the following tools: Conda, Singularity, Docker. In addition, data brokering using athENA requires a prior registration to ENA to get a broker account. athENA is currently used routinely by the SeBiMER, to publish sequencing data in behalf of Ifremer research teams. Ongoing developments include procedure to submit individual and Metagenome-Assembled Genome assemblies to ENA.

References

1. Mark D Wilkinson, Michel Dumontier, I Jstrand Jan Aalbersberg. et al. The FAIR Guiding Principles for scientific data management and stewardship. *Sci Data* 3, 160018, 2016.
2. European Nucleotide Archive, webin-cli, GitHub repository, <https://github.com/enasequence/webin-cli>, 2019.
3. Miguel Roncoroni, Bert Drosbeke, Ignacio Eguinoa et al. A SARS-CoV-2 sequence submission tool for the European Nucleotide Archive (Z. Lu, Ed.). *Bioinformatics*. 10.1093/bioinformatics/btab421, 2021.
4. Frédéric Bigey, *ena-submit*, GitHub repository, <https://github.com/bigey/ena-submit>, 2019.
5. Pauline Auffret, *athena*, Gitlab repository, <https://gitlab.ifremer.fr/bioinfo/workflows/athena>, 2023.

Epitranscriptomic analyses by Nanopore direct RNA sequencing at the I2BC next-generation sequencing facility

Rania Ouazahrou* , Claude Thermes , Erwin Van Dijk , Yan Jaszczyszyn , Abdenacer Mehdi , Celine Hernandez , and Delphine Naquin*¹

¹Institut de Biologie Intégrative de la Cellule – Commissariat à l'énergie atomique et aux énergies alternatives : DRF/I2BC, Université Paris-Saclay, Centre National de la Recherche Scientifique : UMR9198 – Bâtiment 26, 1 avenue de la Terrasse, 91198 Gif/Yvette cedex, France

Résumé

Core facilities aim to give their users access to the newest technologies and scientific methods, which appear and evolve rapidly. Since its creation in 2010, the mission of the I2BC next-generation sequencing facility (PSI2BC) is to provide the scientific community, whether academic or industrial, with services and support in the domain of high throughput sequencing and its applications in functional genomics and transcriptomics. We present an overview of our most recent collaborations for the detection of direct RNA modifications, based on Nanopore Technologie (ONT).

The field of epitranscriptomics has evolved significantly in recent years, as evidenced by the development and publication of numerous modification detection software. Exploiting various aspects of ONT sequencing data, i.e. electrical signal or base calling errors, the use of these tools alone is unfortunately not sufficient to validate the modified positions. For this reason, we conducted, different tests on known data sets before using them in our projects.

Here we present the results of our tests on the different datasets as well as the outcome of two collaborations relating to viral RNA and tRNA sequencing. According to our results, these tools should be used with caution due to the high number of false positives and to the complexity of the effects of RNA modifications on direct RNA nanopore sequencing, and require adequate experimental validation.

Acknowledgements

The I2BC sequencing facility is supported by France Genomique (funded by the French National Program "Investissement d'Avenir" ANR-10-INBS-09).

Mots-Clés: epitranscriptomic, Nanopore, RNA modifications, ONT

*Intervenant

PredomicsApp: R shiny web application for interpretable and accurate construction of prediction models for OMICs data

Gaspar ROY¹, Eugeni BELDA^{1,2}, Youcef SKLAB¹, Edi PRIFTI^{1,2} and Jean-Daniel ZUCKER^{1,2}

¹ IRD, Sorbonne University, UMMISCO, Complex Systems Modelling International Laboratory

² INSERM, Sorbonne University, NUTRIOMICS, Nutrition and Obesities Systemics Approaches Laboratory

Corresponding Author: gaspar.roy@ird.fr, edi.prifti@ird.fr

<https://predomics.ummisco.fr>

The gut microbiome plays an essential role in human health, and machine learning algorithms have been pivotal in the identification of microbial biomarkers associated to different disease conditions. However, most of these algorithms behave as “black-boxes” - results of the predictions are difficult to interpret biologically, while interpretability remains essential for their clinical use in the context of precision medicine. *Predomics [1]* is an R package that combines both state-of-the-art-methods (SOTA) including random forest, GLMNET and support vector machines (SVM) with an ecosystem inspired family of models (Bin/Ter/Ratio models) that provide accurate predictive signatures with unprecedented interpretability. To facilitate their usage for non-computational scientists, a Shiny application has been developed. This application is user-friendly and flexible according to user requirements, allowing to easily upload data, launch experiments and explore results with a simple interface. This application is based on projects created by the user that are data-driven : a project is linked to the data that will be imported in it. This data can then be explored or filtered. Various experiments can then be run with various algorithms and parameters. The app can then automatically build and store interpretable and robust predictive models for microbiome binary classification tasks. This allows users to explore signatures built on different multi-omics spaces with a rich set of graphical panels for result visualization. The best models and their composition are easily accessible, but it is also possible for the user to compare the family of best models, explore their composition or even focus on a single specific model. With PredomicsApp, we hope to grant to bioinformaticians an easy access to powerful interpretable models.

References

1. Prifti, Edi et al. “Interpretable and accurate prediction models for metagenomics data.” GigaScience vol. 9,3 (2020): g1aa010. doi:10.1093/gigascience/g1aa010

The Rhodoexplorer Genome Database: a Multi-Scale Genomic and Transcriptomic Data Resource for the Red Algae

Loraine BRILLET-GUÉGUEN^{1,2}, Tobias WÖRTWEIN⁴, Arthur LE BARS^{3,2}, Agnieszka P. LIPINSKA^{1,4}, Stacy A. KRUEGER-HADFIELD⁵, Olivier GODFROY¹, Simon DITTAMI¹ and Erwan CORRE²

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

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

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

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

⁵ Department of Biology, University of Alabama at Birmingham, 1300 University Blvd, Birmingham, AL, 35294

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

Abstract

Red algae are of interest for a number of aspects, including their life cycles and reproductive biology, as domesticated and cultivated species, as invasive species, and for their characteristic cell walls. The Rhodoexplorer project [1] aims to explore the evolution of biological complexity in the red algae through the establishment of a multi-scale genomic data resource for the red algae, including public data and new sequenced and annotated genomes. The Rhodoexplorer project involves an international consortium including partners at the Max Planck Institute (MPI) in Tübingen (Germany), the Roscoff Biological Station (France), the University of Sao Paulo (Brazil), Universidad Austral de Chile (Chile), the University of Alabama at Birmingham (USA), the University of Charleston (USA), GEOMAR, Kiel (Germany) and the University of Oldenburg (Germany).

The ABiMS platform has developed a new web portal (<https://rhodoexplorer.sb-roscoff.fr>) to house the annotated genome sequences and associated resources, including genome browsers; information about the sequenced strains; assembly and annotation metrics; data download facilities; and BLAST facilities. In the context of other local genome portal projects, and of the European Reference Genome Atlas project, in partnership with two other Breton bioinformatics platforms, we are setting up an automated, modular, and FAIR system for the provision, visualization and processing of large-scale genome data: the BEAURIS pipeline (<https://gitlab.com/beauris/beauris>) [2].

References

1. The rhodoexplorer platform for red algal genomics and whole genome assemblies for several gracilaria species. Agnieszka P. Lipinska, Stacy A. Krueger-Hadfield, Olivier Godfroy, Simon Dittami, et al. bioRxiv 2023.03.20.533491; doi: 10.1101/2023.03.20.533491
2. BEAURIS: an automated, modular, and FAIR system for large-scale genome data management. Matéo Boudet, Loraine Brillet-Guéguen, Arthur Le Bars, Karine Massau, Laura Leroi, Alexandre Cormier, Patrick Durand, Erwan Corre, Anthony Bretaudeau. Talk, JOBIM 2023.

The Migale bioinformatics core facility

Valentin LOUX^{1,2}, Mouhamadou BA^{1,2}, Christelle HENNEQUET-ANTIER^{1,2}, Mahendra MARIADASSOU^{1,2}, Véronique MARTIN^{1,2}, Cédric MIDOUX^{1,2,3}, Olivier RUÉ^{1,2}, Valérie VIDAL^{1,2} and Sophie SCHBATH^{1,2}

¹ Université Paris-Saclay, INRAE, MaIAGE, Domaine de Vilvert, 78350, Jouy-en-Josas, France.

² Université Paris-Saclay, INRAE, BioinfOmics, MIGALE bioinformatics facility, Domaine de Vilvert, 78350, Jouy-en-Josas, France.

³ Université Paris-Saclay, INRAE, PRocédés biOtechnologiques au Service de l'Environnement, 1 rue Pierre-Gilles de Gennes, CS10030, 92761, Antony, France.

Corresponding Author: valentin.loux@inrae.fr

The Migale bioinformatics facility is a team of INRAE's MaIAGE research unit (Applied Mathematics and Computer Science, from Genome to the Environment). It has been providing services to the life sciences community since 2003.

Migale is an open platform, that offers four types of services ;

- an open infrastructure dedicated to life sciences data processing,
- dissemination of expertise in bioinformatics,
- design and development of bioinformatics applications,
- text mining, genomic, metagenomic and metatranscriptomic analysis.

Migale is part of the French Institute of Bioinformatics (IFB) and France Génomique projects. It has an ISO9001 certification and is also one of the four platforms which compose BioinfOmics, the national Research Infrastructure in bioinformatics of INRAE.

The poster will illustrate the platform services with examples chosen from various projects achieved this year or having recently started such as "Grand Défi Ferments du Futur", in which Migale coordinates the development of a data warehouse to gather all information on microorganisms and microbial consortia involved in food fermentation.

A complete description of Migale facility's service offer is available on its website : <https://migale.inrae.fr>

L'institut Français de Bioinformatique: Centre de Référence Thématique Biologie-Santé dans l'écosystème Recherche Data Gouv

Jacques VAN HELDEN¹, Hélène CHIAPELLO², Olivier COLLIN³, Thomas DENECKER¹, Marie-Dominique DEVIGNES⁴, Jean-François DUFAYARD⁵, Alban GAINARD⁶, Nadia GOUE⁷, Frédéric DE LAMOTTE⁸, Julien SEILER¹ et le groupe de travail 'MISSION SCIENCE OUVERTE ET INTEROPERABILITE' DE L'IFB.

¹IFB-Core CNRS, 91057, Evry, France ; ²MaIAGE INRAE, 78350, Jouy-en-Josas, France ; ³Plateforme GenOuest IRISA-INRIA, 35042, Rennes, France ; ⁴Université de Lorraine CNRS Inria LORIA, 54000 Nancy, France ; ⁵CIRAD, 34398, Montpellier Cedex5, France ; ⁶Plateforme BIRD Institut du thorax, 44007 Nantes, France ; ⁷Plateforme AuBi Mésocentre Clermont Auvergne, 63178 Aubière Cedex, France, ⁸Département de Biologie et Amélioration des Plantes INRAE, 34398, Montpellier Cedex5, France.

Corresponding Author: Jacques.van-Helden@france-bioinformatique.fr

L'écosystème [Recherche Data Gouv](#) a été pensé et créé par le Ministère de l'Enseignement Supérieur et de la Recherche comme un dispositif d'accompagnement de la communauté scientifique pour le partage et l'ouverture des données de la recherche. Il s'agit de soutenir les équipes de recherche dans leur travail de structuration des données pour les rendre conformes aux principes "FAIR" : Faciles à trouver, Accessibles, Interopérables, Réutilisables. Cet écosystème comporte plusieurs composantes : trois modules d'accompagnement des équipes de recherche : [ateliers de la donnée, centres de référence thématiques, et centres de ressources](#), et deux modules pour déposer, publier et signaler des données : [un entrepôt de données national](#) et un **catalogue** pour rechercher les données publiées sur l'entrepôt ou sur des entrepôts externes.

En 2022, l'[Institut Français de Bioinformatique](#) (IFB) a été sollicité pour constituer le premier Centre Référence Thématique (CRT) dans le domaine de la biologie et de la santé. Cette notification s'accompagne de deux objectifs principaux. 1) Animation de l'ensemble des communautés biologie-santé, afin de définir les standards de FAIRisation : métadonnées minimales, vocabulaires contrôlés ou ontologies, entrepôts recommandés, etc. ainsi que mise à disposition d'un référentiel de bonnes pratiques et de formations en science ouverte dans les domaines biologie-santé. 2) Développement ou recommandation d'outils stratégiques pour aider à gérer les données et leurs métadonnées tout au long de leur cycle de vie, en proposant des exemples d'utilisation et des synergies d'usage.

Pour répondre au premier objectif, l'IFB a suscité l'émergence d'un groupe de travail "Données de la Recherche" au sein de l'ensemble des Infrastructures Nationales en Biologie et Santé (club des INBS). Ce groupe travaillera à construire un panorama des types de données de recherche en biologie-santé, qui sera étayé par la collecte des standards et normes utilisés ou recommandés par les communautés scientifiques représentées dans les INBS. L'étendue des bonnes pratiques ainsi recensées dépendra de l'avancée de chaque domaine dans la mise en œuvre des principes FAIR et de la science ouverte. Par ailleurs, l'IFB est déjà impliqué dans plusieurs formations relatives à la FAIRisation des données, qui feront partie du référentiel de bonnes pratiques et de formations élaborées par le CRT biologie-santé (en lien avec le centre de ressources [DoRANum](#)).

En ce qui concerne le deuxième objectif, l'IFB s'appuie sur un ensemble de développements initiés en interne et dans le cadre de projets de recherche (MuDIS4LS, CONVERGE). Quelques exemples : l'outil [FAIR-Checker](#) d'évaluation automatique de pages web pour le respect des principes FAIR, la spécialisation des plans de gestion de données (PGD) sous forme de modèles propres à telle ou telle infrastructure (en lien avec le centre de ressources [OPiDoR](#)), la valorisation des PGD en les rendant actionnables par des programmes, un outil de courtage de données (METARK) pour faciliter et fluidifier des dépôts de données et métadonnées FAIR dans les entrepôts nationaux ou internationaux, le tableau de bord OpenLink qui propose une vue d'ensemble des données et outils associés à chaque projet de recherche, depuis le PGD jusqu'aux entrepôts, en passant par le cahier de laboratoire électronique et l'ensemble des solutions de stockage utilisés couramment par les chercheurs. Ces outils seront mis à disposition des équipes de recherche, ainsi que possiblement d'autres outils existants déjà adoptés par certaines communautés.

L'invitation à constituer le premier CRT biologie-santé dans l'écosystème Recherche Data Gouv est une reconnaissance de l'implication de l'IFB et de l'ensemble de ses plateformes pour la science ouverte, notamment à travers les formations et les projets. La position de l'IFB comme nœud français du réseau européen des plateformes de bioinformatique [ELIXIR](#) lui permettra aussi de contribuer à la dimension européenne du CRT biologie-santé, appelé à se positionner au sein de la coordination européenne pour la science ouverte ([EOSC](#)).

RGB: the Guadeloupean Network of CRBs

Stanie Gaete*¹, Damien Meyer , Michel Naves , Nilda Paulo De La Reberdiere* ,
Dominique Dessauw , Yoanna Faure , Marie Umber , and Jacqueline Deloumeaux

¹CHU de la Guadeloupe – Guadeloupe

Résumé

Guadeloupe is the only French overseas territory with biological resource centers (BRCs) representing 4 different kingdoms (Micro-organisms and Vectors, Plants, Animals and Humans). Initiated as early as 2006, these BRCs have federated around the project to create the Guadeloupe BRC Network (GBR) proposed by Karubiotec™ in 2021. This network aims to improve the visibility of local BRCs in the research projects carried out thanks to the biological resource collections they make available. It is the guarantee of the biodiversity of the island's genetic heritage and allows Guadeloupean researchers to be a force of proposal in the submission of multi-thematic projects. It is consistent with the national desire to develop biobanks (Health Innovation Plan 2030) and with the "One Health" research axis (ERDF PO 2021-2027). Communication on the activities of local BRCs to professionals, scientists, the educational world and the public is essential for a better understanding of the world of BRCs, international research issues and interactions between the different environments. The emergence of new infections in various kingdoms recently demonstrates the need for a global approach. The RGB has essential assets for increasing the awareness of research consortia to multi-thematic approaches based on these structuring tools, which have technical know-how and scientific expertise, combined with ethical and regulatory knowledge. The professionalization of Guadeloupe's BRCs is a proof of quality and a criterion of respect for the ethics of the heritage of which they are guardians.

Mots-Clés: Biological Cessource Centers, Karubiotec, Mivec, Carare, Tropical plants, earthly kingdom.

*Intervenant

MERIT : réseau MetiER en bioinformaTique

Audrey BIHOUEE¹, Hélène CHIAPELLO^{2,7}, Erwan CORRE^{3,7}, Vincent LEFORT^{4,7}, Alexandra LOUIS⁵, Jimmy VANDEL⁶.

¹ BiRD bioinformatics platform, Université de Nantes

² MaIAGE team, INRAE, Jouy en Josas - Paris

³ ABiMS bioinformatics platform, CNRS - INEE, Roscoff

⁴ ATGC bioinformatics platform, LIRMM UMR5506, CNRS - INS2I, Université de Montpellier

⁵ Dyogen team, IBENS, CNRS - INSB, Paris

⁶ Bilille bioinformatics platform, PLBS, CNRS - INSB, Lille

⁷ Institut Français de Bioinformatique, CNRS UMS 3601, France

Corresponding Author: vincent.lefort@lirmm.fr

L'évolution des technologies d'acquisition de données en sciences du vivant génère une avalanche de données. Les métiers permettant de gérer, traiter et analyser ces données représentent aujourd'hui un enjeu primordial. Le métier de bioinformaticien-ne est intrinsèquement interdisciplinaire. Les ingénieur-e-s du domaine dépendent de la Branche d'Activité Professionnelle des sciences du vivant, de la terre et de l'environnement (BAP A) ou de celle de l'informatique, statistiques et calcul scientifique (BAP E) mais les activités des bioinformaticien-ne-s peuvent emprunter une grande diversité d'emploi-types qui reflète le contexte interdisciplinaire dans lequel le métier s'exerce et les nombreuses thématiques scientifiques abordées.

La communauté doit faire face à l'évolution technologique rapide, les ingénieur-e-s bioinformaticien-ne-s ont régulièrement besoin de se former. Par ailleurs, pour rompre leur isolement thématique, les ingénieur-e-s bioinformaticien-ne-s ont besoin de lieux d'échanges leur permettant de se rencontrer et de partager leurs expertises.

Nous sommes donc heureux de vous annoncer la création d'un réseau MétiER en bioinformaTique (MERIT: <https://merit.cnrs.fr/>). Ce réseau est ouvert à l'ensemble des personnels, de toutes les tutelles, intéressés par les animations métier en ingénierie bioinformatique (ingénieur-e-s, chercheur-euse-s). Le réseau MERIT a pour but de fédérer cette communauté afin de participer au maintien et au développement des compétences et de limiter l'isolement professionnel. Les actions proposées ont pour objectifs d'assurer une veille technologique grâce à la création de groupes de travail, d'organiser des formations et de travailler à la reconnaissance du métier de bioinformaticien-e et des problématiques liées à l'interdisciplinarité.

Ce nouveau réseau métier s'est créé grâce à l'INSB et l'INS2I, et avec le soutien de l'INEE, l'INC, la SFBI et l'IFB. Les activités du réseau sont envisagées en concertation avec les autres acteurs nationaux ([IFB](#), [AVIESAN](#),...). Les actions communes en partenariat sont privilégiées.

Ferments du Futur : une plateforme unique en Europe qui entend accélérer la recherche et l'innovation sur les ferments, les aliments fermentés et la biopréservation dans les 10 prochaines années

Sophie SCHBATH^{1,2}, Valentin Loux^{1,2}, Madeleine Spatz³, Damien Paineau³ et l'équipe Ferments du Futur

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

² INRAE, BioinfOmics, MIGALE bioinformatics facility, 78350, Jouy-en-Josas, France

³ Ferments du Futur, INRAE Transfert, 75015, Paris, France

Corresponding Author: Sophie.Schbath@inrae.fr

Avec l'augmentation des risques environnementaux et climatiques, les attentes sociétales en termes d'alimentation ont radicalement changé vers une alimentation plus sûre, plus saine, plus durable. Les modes de consommation ont évolué vers une baisse de la consommation de protéines animales en faveur de protéines végétales et des exigences plus aigües des consommateurs (plus de naturalité, de durabilité, mais également de goût, de plaisir, de bien-être, voire de bénéfique santé). Les aliments et boissons fermentés, par l'action de champignons et/ou bactéries, permettent de transformer la matière première et d'apporter de nouvelles propriétés organoleptiques et/ou de conservation. De plus, ils sont une source de micro-organismes vivants et peuvent être un levier dans la modulation de l'écosystème environnement/hôte/microbiote intestinal. C'est dans ce contexte que s'est lancé à l'automne 2022 Ferments du Futur (FF), soutenu à hauteur de 48,3 M€ par France 2030 [1]. FF entend accélérer dans les 10 ans à venir la recherche et l'innovation sur les ferments, les aliments fermentés et la biopréservation.

Pour se faire, FF a réuni un écosystème paritaire public/privé, composé à ce jour de 37 partenaires réunissant des établissements d'enseignement supérieur et de recherche et des partenaires industriels dans le domaine des ferments et des aliments fermentés. Cet écosystème va s'appuyer sur des capacités de pointe pour lever les verrous scientifiques et technologiques, grâce à une entité unique, distribuée sur plusieurs sites : une plateforme technologique distribuée au sein de sept unités de recherche INRAE (MaIAGE, MGP, Micalis, SayFood, SPO, STLO, UMR) spécialisées en microbiologie, procédés, bioinformatique et une plateforme d'innovation installée sur la plateau de Saclay et dotée de plateaux de criblage, fermentation et prototypage d'aliments fermentés. En s'appuyant sur l'Intelligence Artificielle et la Science des données, FF vise à développer des ferments et des produits fermentés innovants, répondant à des fonctionnalités précises, à travers la conception rationnelle de consortia microbiens issus de la biodiversité naturelle et des procédés de fermentation optimisés. Centré sur l'alimentation humaine, FF pourra s'étendre progressivement vers d'autres secteurs comme l'alimentation animale, la santé, l'agriculture ou l'environnement.

Cette plateforme, unique en Europe, grandira progressivement pendant l'année 2023. Tous les laboratoires publics français qui souhaitent participer à cette dynamique et à cet écosystème peuvent candidater aux appels à projets précompétitifs annuels. FF entend également développer des formations pour accompagner le développement des capacités de fermentation, notamment au niveau industriel.

La plateforme bioinformatique Migale coordonne le développement informatique de l'entrepôt de données de FF visant à rassembler toutes les informations sur les micro-organismes et les consortia microbiens pouvant intervenir dans la fermentation des aliments. Ces données sont de nature variée --- omiques, phénotypiques, capacités métaboliques, etc. --- et alimenteront les approches prédictives développées dans le projet pour la conception de nouveaux aliments fermentés.

Acknowledgements

Ce travail a bénéficié d'une aide de l'État gérée par l'Agence Nationale de la Recherche au titre de France 2030 portant la référence ANR-22-GDFF-0001

References

1. <https://www.inrae.fr/actualites/innover-lalimentation-demain-lancement-operationnel-du-grand-defi-ferments-du-futur>

Building Research Capacity for Diagnostics, Exposome, and Bioinformatics in the Caribbean: A Collaborative Thesis Project

Jean ARNAUD¹, Timothy FORD¹ and David COUVIN²

¹ Ford Lab, 3 Solomont Way #005-#007, Lowell, MA, 01854, USA

² Unité Transmission, réservoir et diversité des pathogènes, Institut Pasteur de la Guadeloupe, Les Abymes, Guadeloupe, France

Corresponding Author:

Jean_Arnaud@uml.edu, Jean.chem.Arnaud@gmail.com, jarnaud2009@gmail.com

ABSTRACT

To what extent does the Lived Environment exacerbate the prevalence of certain diseases?

This research project seeks to investigate the impact of the lived environment exposome on the prevalence of certain diseases in the Caribbean region. The primary objective is to bridge the research capacity gap in diagnostics, exposome, and bioinformatics specific to the Caribbean. Through a participatory research and action methodology, which features diverse means of communication (including blogs, WhatsApp group, etc) the project will recruit and train medical students and health professionals. It will also create opportunities for exchanges and discussions on selected topics, facilitating the development of research and focus groups in areas of regional and personal interests.

The project aims to compile the research findings into a white paper, along with other publication opportunities. By doing so, it aims to raise awareness and attract investment for further research and capacity building in the Caribbean. The goal is to address the research capacity gap effectively and enhance understanding and capabilities to propose measures to mitigate these health factors.

In summary, this research project focuses on exploring the extent to which the exposome exacerbates the prevalence of specific diseases in the Caribbean. Through recruitment, training, and participatory research, the project aims to bridge the research capacity gap and generate valuable insights for mitigating these health factors. The findings will be compiled into a white paper and other publications to promote awareness and attract investment, ultimately fostering further research and capacity building in the Caribbean region.

Executive Summary

1. Introduction and Objective:

This executive summary details a research project that aims to investigate the impact of the lived environment exposome on the prevalence of certain disease in the Caribbean, while also addressing research gap capacity and scientist representation from the region, specifically within the context of computational biology, exposome analysis and diagnostics.

2. Participatory Research Methodology:

Through participatory efforts, the project will engage health practitioners to participate in dialogue, literature reviews, surveys, and questionnaires to help identify topics of concerns for focused activities. Blogs, Interviews and other materials will be communicated through using diverse communication tools, such as Twitter, WhatsApp and Facebook. Ultimately, diverse set of focus areas of inquiry will emerge. The project seeks to encourage open science publication and public discussions.

3. Compilation of Research Findings:

The project aims to compile the research findings into a comprehensive white paper, supplemented by other publication opportunities. By disseminating the results, the project seeks to raise awareness about the significant impact of the exposome on disease prevalence in the Caribbean. Additionally, it is hoped that these efforts can attract investment for further research and capacity building efforts in the region.

4. Addressing the Research Capacity Gap:

The overarching goal of this research project is to effectively address the research capacity gap and enhance understanding and capabilities in diagnosing and mitigating the adverse health factors associated with the lived environment exposome. By leveraging recruitment, training, and participatory research, the project aims to generate valuable insights that can inform measures and interventions aimed at reducing disease prevalence in the Caribbean.

5. Stakeholder Engagement and Awareness:

Through the compilation of research findings, the project aims to raise awareness among various stakeholders, including researchers, policymakers, and the general public.

6. Conclusion and Overall Impact:

In conclusion, this research project focuses on investigating the extent to which the lived environment exposome exacerbates the prevalence of specific diseases in the Caribbean. By actively involving diverse stakeholders, bridging the research capacity gap, and compiling research findings into accessible publications, the project aims to promote awareness, attract investment, and foster continued research and capacity building efforts in the Caribbean to mitigate the adverse health factors associated with the lived environment.

List of Participants

- Abraham, Anne-Laure, 218
Achaz, Guillaume, 42
Adam, Catherine, 251
Ahraoui, Amèle, **205**
Ait-El-Mkadem Saadi, Samira, 203
Aite, Meziane, 158, 249
Alves, Isabel, 151
Amar, Patrick, 248
Amaya, Diego, **243**
Amblard, Elise, **202**, 211
Antunes, Mathias
Arduin, Hélène, 38
Argiro, Laurent, 230
Armand, Florence
Armant, Olivier, 146
Arnaud, Jean, **264**, **292**
Arquet, Rémy, 162
Arzul, Isabelle, 272
Aschard, Hugues, 160
Aubé, Johanne, **216**
Aubert, Grégoire, 212
Auboeuf, Didier, 140
Aubourg, Nicolas
Aubry, Marc, 113
Audit, Benjamin, 172
Audrézet, Marie-Pierre, 239
Auer, Lucas
Auffret, Pauline, 68, **283**
Aury, Jean-Marc, 149
- Ba, Mouhamadou, 287
Bacq, D., 151
Badis, Yacine, 224
Bailly-Bechet, Marc, 149
Bailly-Chouriberry, Ludovic, 196
Bal, Antonin, 232
Ballandras-Colas, Allison, **14**
Bambou, Jean-Christophe, **65**, 101, 162
Bannwarth, Sylvie, 203
Baptista, Anthony, 159, 210
Barbe, Valérie, 220
Barbier, Jérémy, 172
Bardon, Alizée, 68, 283
Bardou, Philippe, 236
Barloy, Dominique, 173
Barloy-Hubler, Frédérique, 173
Barnabé, Agnès, **196**
- Barth, Dominique, 36
Bastide, Paul, 90
Bastien, Sylvère
Batantou, Degrâce, **21**
Battistel, Clémentine, **272**
Baucheron, Sylvie, 218
Baudot, Anaïs, **15**, 159, 210, 230, 252, 281
Beaudoin, J.-C., 151
Beaumont, Guillaume
Bécavin, Christophe
Becht, Etienne, 72
Béesau, Julie, 257
Belcour, Arnaud, 158, 249
Belda, Eugeni, 285
Bellengier, Juliette
Bellot, Clément
Belser, Caroline, 149
Ben Khedher, Mariem
Ben-Jemaa, Slim, 236
Benchouaia, Médine
Bennour, Juba
Benoit, Clément
Beramice, David, 101
Berthelier, Charlotte
Bertrand, Vadim, 202, **211**
Besse, Florence, 247
Bessoltane, Nadia
Besson, Aurore
Beust, Cécile, **159**
Bezault, Etienne, **217**
Bielle, Franck, 197
Bihouée, Audrey, **290**
Bioteau, Audrey
Bittner, Lucie, 235
Bize, Ariane, 282
Blanchard, Yannick, 180
Blanché, Hélène, 151
Blanquart, Samuel, 158, 249
Blin, Aurélie, 186
Blot, Lauren, 247
Blum, Yuna, 113
Bocaly, Méliissa, 225
Bohers, Elodie, 191
Boland, A., 151
Bon, Céline, 154
Bonis, Mathilde, 155

Bonneau, Mathieu, **81**
 Bonzom, Jean-Marc, 146
 Bordenave, Julie, 38
 Bottini, Silvia, 83, 203
 Bouallegue, Syrine
 Bouamout, Hajar, **190**
 Bouanich, Andréa, **260**
 Boubred, Djamel
 Boudet, Matéo, **136**
 Boulaimen, Youssef
 Boulanger, Marine, 179
 Boulinguez, Matthieu, **268**
 Bourbeillon, Julie, 260
 Boussaha, Mekki, 236
 Boutegrabet, Warda
 Boutonnat, Jean, 209
 Bouvier, Guillaume, 274
 Boyer, Théophile, **232**
 Bozorgan, Anne, 233
 Branger, Maxime, 199
 Bray, Frédéric
 Breard, Emmanuel, 180
 Bresso, Emmanuel
 Bretaudeau, Anthony, 136, 176, 270
 Breurec, Sebastien, **50**
 Briandet, Romain, 179
 Bridier, Arnaud, 179
 Briere, Galadriel, **210**
 Brière, Marie-Galadriel
 Briere, Marie-Galadriel, 159
 Brillet-Guéguen, Lorraine, 136, 184, 276, **286**
 Bringel, Françoise, 237
 Bročić, Jovana
 Brocic, Jovana, **197**
 Brunet, Theo
 Bruno, Morgane
 Bui quynh, Trang, 182
 Burner, Fred, 225
 Burstin, Judith, 212

 Cabanac, Sébastien, **192**
 Caderhoussin, Alann, 162
 Calcagno, Vincent, 176
 Calenge, Fanny, 218
 Calia, Giulia, **83**
 Callon, Cécile, 220
 Calvez, Elodie, **23**
 Cambon-Bonavita, Marie-Anne, 216
 Camenen, Etienne
 Canier, Lydie, 272
 Cariou, Marie
 Carpentier, Marie-Christine
 Carpentier, Mathilde, 235
 Casse, Nathalie, 221
 Castellani maria, Aparecida, 183
 Castro Alvarez, Javier, 233
 Cavaiuolo, Marina, 155
 Cawley, Adam, 196
 Cazau, Dorian, 257

 Cazenave, Damien, **188**
 Cedilnik, Nicolas, 205
 Cestaro, Alessandro, 83
 Chaffron, Samuel, **28**, 281
 Chalvignac, Richard, 232
 Chantalat, Sophie, 244
 Charron, Raphaël, 179
 Chassagnol, Bastien, **72**
 Chatelain, Estelle
 Chauvot De Beauchêne, Isaure, **242**
 Checa Ruano, Luis, 274
 Chen, Fati, 277
 Chénais, Benoît, 221
 Cherchame, Emeline, **109**
 Chesnais, Virginie, 155, **161**, 180
 Chevalier, Céline, **230**
 Chevignon, Germain, 272
 Chiapello, Hélène, 281, 282, 288, 290
 Chikhi, Rayan, **16**
 Choquet, Caroline, 230
 Chouali, Hanae
 Choury, Alice, 101
 Chuffart, Florent, 209
 Clement, Julie, 283
 Cock, Mark, 184
 Cognat, Valérie
 Coignard, Bruno, 233
 Collen, Jonas, 158, 249
 Collin, Olivier, 288
 Coluzzi, Charles, **42**
 Condamine, Bénédicte
 Confais, Johann, 174
 Corcos, Laurent, 231
 Cordonnier, Sébastien, 225
 Corler, Enora
 Cormier, Alexandre, 68, 136, 184, **201**, 262, 283
 Cornec-Le Gall, Emilie, 239
 Corporeau, Charlotte, 262
 Corre, Emma, **178**
 Corre, Erwan, 136, 184, **276**, 286, 290
 Corre, Sebastien, 113
 Couture, Carole
 Couvin, David, **66**, 162, 193, **251**
 Croce, Olivier
 Cruaud, Corinne, 220
 Cueff-Gauchard, Valérie, 216
 Cunnac, Sébastien, 200
 Curien, Gilles

 Da Rocha, Martine, 149, 186, 259
 Dablanc, Axel
 Dadvisard, Maylis
 Dallet, Romain
 Dameron, Olivier, 281
 Danchin etienne, GJ, 259
 Danchin, Etienne, 83, 149
 Dantec, Laurent, 101
 Darbo, Elodie
 Daric, Vladimir

Darnige, Eden
 Davy, Martin, **248**
 De Beauchene, Isaure
 De Brevern, Alexandre, **37**
 De Lahondès, Raynald, 219
 De Lamotte, Frédéric, 288
 De Massy, Bernard, 283
 De Sousa Violante, Madeleine
 Decroocq, Veronique, 182
 Dedeine, Franck, 134, 226
 Dekeyser, Thibault, 132
 Delage, Ludovic, 158, 249
 Delaunay, Stéphane, 250
 Delbès, Céline, 220
 Delcourt, Vivian, 196
 Deleury, Emeline, 186
 Deleuze j., F., 151
 Delfau-Larue, Marie-Hélène, 191
 Deloumeaux, Jacqueline, 289
 Deman, Vincent
 Demoulin, Nathalie, 239
 Denecker, Thomas, 282, 288
 Denise, Alain, 36
 Denni, Sukanya, **182**
 Dereeper, Alexis, 162, 200
 Derian, Quentin
 Derouin, Margot, **208**
 Dérozier, Sandra
 Despot-Slade, Evelin, 149
 Dessauw, Dominique, 289
 Destras, Gregory, 232
 Devarieux, Oriane, 101
 Devignes, Marie-Dominique, 243, **288**
 Devriese, Magali, 243
 Diamant, Anna
 Dina, Christian, 151
 Dittami, Simon, 158, 224, 249, 286
 Djaout, El Hacene
 Djebali, Sarah, 160, 240
 Domagala, Marcin, 38
 Donati, Claudio, 83
 Doré, Guillaume, **173**
 Doublet, Benoît, 218
 Drula, Elodie, 268
 Dufayard, Jean-François, 288
 Dugat-Bony, Éric, 220
 Dugourd, Aurélien, **29**
 Dulary, Cécile, 244
 Dumont, Florent, **278**
 Dunand, Christophe, 192
 Duplessis, Sébastien, 178
 Dupont, Mathieu, 257
 Durand, Lucile, 216
 Durand, Patrick, **68, 194**, 201, 272, 283
 Duvaux, Ludovic, 182

 Eid, Jad, **142**
 El Assimi,
 El Ghaziri, Angelina, 260

 El Khatib, Mariam, **176**
 El-Hami, Loubna, 203
 Eppe, Gauthier, 251
 Estoup, Arnaud, 90
 Euphrasie-Clotilde, Lovely, **195**
 Eveillard, Damien

 Fall, Ahmad, 206
 Fall, Mame Seynabou
 Fall, Sidy
 Faure, Roland, **124**
 Faure, Yoanna, 289
 Félicité, Yoann, 101
 Ferdinand, Séverine, **162**
 Ferrane, Assia, 151
 Feudjio, Olivier, **256**
 Feuillet, Dalila, 101
 Fierville, Morgane
 Fiévet, Ghislain, **204**
 Fin, B., 151
 Fiston-Lavier, Anna-Sophie
 Fisun, Michel, 160
 Fleury, Elodie
 Flisch, Alan, 233
 Flores, Raphaël-Gauthier, 212
 Flot, Jean-Francois, 124
 Fodil, Mostefa, 221
 Fontanille, Emmanuelle
 Fontrodona, Nicolas, 140
 Fortuné, Antoine
 Foulon, Sidonie, 208
 Fourcot, Marie
 Fournié, Jean-Jacques, 38
 Fournier, Frantz, 250
 Fraboulet, Rose-Marie, **113**
 François, Olivier, 147
 Frioux, Clémence, 158, 249
 Frouin, Eléonore
 Fumey, Julien

 Gaete, Stanie, **289**
 Gaignard, Alban, 281, 288
 Galan, Clément, **175**
 Galibert, Marie-Dominique, 113
 Gamiette, Gélixa, **148**
 Ganofsky, Jérémy
 Gansevoort, Ron, 239
 Garali Zinedine, Imène, 146
 Garcia - Van Smévoorde, Margot
 Garcia, Céline, 272
 Garcia, Damien, **245**
 Garcia, Patrice, 196
 Garcia-Van Smévoorde, Margot, **24**
 Gardon, Hélène, **220**
 Gareau, Thomas, 109
 Gaspar, Nathalie, 150
 Gaspin, Christine
 Gautier, Mathieu, 186
 Gautier, Romain

Gavory, Frédérick, 220
 Gazengel, Kévin, **263**
 Gendre, Julia, 220
 Genete, Mathieu
 Génin, Emmanuelle, 132, 151
 Genkyst Study Group, The, 239
 Gennesseaux, Laureline, **157**
 Geoffroy, Véronique, **111**
 Gerber, Zoé, **160**
 Ghata, Adeline, 230
 Gianfrotta, Coline, **36**
 Gibert, Audrey, 140
 Glaser, Philippe, 42, 162
 Glibert, Florence, 244
 Gobert, Guillaume
 Godfroid, Maxime, 42
 Godfroy, Olivier, 286
 Gomes, Lucie, **191**
 Gondard, Mathilde, 180
 Gonnet, Iphigénie
 Gonzalez-Rizzo, Silvina, 222
 Got, Jeanne, 158, 249
 Gottis, Benjamin
 Goudenegge, David
 Goué, Nadia, 288
 Goujon, Elen, **146**
 Gourdine, Jean-Luc, 101
 Graindorge, Stéphanie
 Gressin, L., 151
 Gros-Martial, Anatole, 257
 Gruel, Gaëlle, 162, **253**
 Guedon, Emmanuel, 250
 Guégan, Justine, 109, 197
 Guéganton, Marion, 216
 Guéguen, Lorraine
 Guermeur, Yann, 242
 Guiavarc'h, Yann, 157
 Guichard, Lucile, 177
 Guignard, Thomas, 111
 Guigon, Isabelle
 Guillemet, Martin, 42
 Guillemot, Vincent
 Guinchard, Sarah
 Guivarch, Mael
 Guivarch, Maël, **151**
 Guyomar, Cervin, 240
 Guyomard, Stéphanie, 251
 György, Beáta, 109

 Hak, Fiona
 Halleger, Martina, 247
 Halluin, Sidonie
 Harris, Peter, 239
 Hassanaly-Goulamhousen, Rahim, 149
 Hayer, Juliette
 Héligon, Christophe, 207
 Hellec, Elisabeth, **262**
 Hennecart, Baptiste, **219**
 Hennechart, Solweig

 Hennequet-Antier, Christelle, 287
 Henriques, Emma
 Henry, Nicolas
 Hergalant, Sébastien, **39**, 204
 Hernandez, Céline, 284
 Herrmann, Carl, **17**, 281
 Hervé, Vincent, **134**, 220, 226
 Herzig, Anthony, **132**, 151
 Hilliou, Frédérique, 176, 183
 Hilpert, Cécile
 Hitte, Christophe
 Hoebeke, Mark, 276
 Hoede, Claire
 Houée, Paméla, 179
 Hubstenberger, Arnaud, 247
 Humeau, Catherine, 157
 Hunault, César, **246**

 Imbert, Baptiste, **212**
 Irlinger, Françoise, 220

 Jacquot, Maude, 272
 Jamilloux, Véronique, 282
 Jardin, Fabrice, 191
 Jardin, Hugo
 Jarrige, Domitille, **237**
 Jaszczyszyn, Yan, 284
 Jay, Flora, 154
 Jean-Luis, Patrick, 222
 Jerome, Emilie
 Joaquim Justo, Célia, 251
 Jossaud, Fabien, **209**
 Josset, Laurence, 232, 233
 Junker, Romane

 Karami, Yasaman
 Khamvongsa-Charbonnier, Lucie, **281**
 Khneisser, Pierre, 150
 Klopp, Christophe, 221
 Knebelmann, Bertrand, 239
 Koebnik, Ralf, 200
 Kon-Kam-Kim, Guillaume, 218
 Kon-Sun-Tack, Fabien, **231**
 Koutsovoulos georgios, D, 149
 Kozlowski, Djampa, 83, 149
 Kravchenko, Anna
 Kreplak, Jonathan, 212
 Kress, Arnaud, 111
 Krueger-Hadfield, Stacy, 286
 Kuchl, Claire
 Kwamou Ngaha, Sandy Frank
 Kwamou Ngaha, Sandy Frank, 219

 Labadie, Karine, 149
 Labory, Justine, **203**
 Lacomblez, Claire
 Lacroix, Thomas
 Lagadec, Ronan, 177
 Laghrissi, Hiba, **247**

Lambourdiere, Josie, 225
 Lamouche, Jean-Baptiste, 111
 Landès, Claudine, 260
 Lapalu, Nicolas, 182
 Lependry, Audrey, **140**
 Lavenier, Dominique, 124
 Le Bars, Arthur, 136, 233, 286
 Le Behec, Antony, 111
 Le Bideau, Gwendal, 203
 Le Corguillé, Gildas
 Le Folgoc, Gaëlle, 151
 Le Mailloux, Guillaume, **90**
 Le Meur, Yannick, 239
 Le Nézet, Louis
 Le Scanf, Enora, 231
 Leblanc, Catherine, 158, 224, 249
 Lebreton, Annie
 Lebrigand, Kevin
 Leclercq, Sébastien, **199**, 218
 Lecoœur, Alexandre
 Lecorguille, Gildas, 276
 Lefeuvre, Maël, **154**
 Lefort, Gaëlle
 Lefort, Vincent, 290
 Legeai, Fabrice, 270
 Legeay, Erwan, 224
 Legras, Mia
 Lemée, Pierre, **179**
 Lemoine, Hugo, **239**
 Lence, Alex, **206**
 Leroi, Laura, 68, 136, 272, 283
 Lerond, Julie, 197
 Lescroart, Fabienne, 230
 Lesongeur, Françoise, 216
 Lespinet, Olivier, 36
 Letexier, Mélanie, **280**
 Leutenegger, Anne-Louise, 208
 Lhotte, Romain, 243
 Lina, Bruno, 233
 Linard, Benjamin, 190
 Lipinska agnieszka, P., 286
 Littner, Eloi
 Liu, Xi
 Loire, Benjamin, **257**
 Lombaert, Eric, 186
 Lopez, Pascal-Jean, 225
 Lopez-Roques, Céline, 259
 Lorrain, Cécile, 178
 Louis, Alexandra, 290
 Loup, Benoît, 196
 Loux, Valentin, 220, **287**, 291
 Ludwig, Thomas, 151
 Luo, Yufei, 72

 Macé, Sabrina
 Mackle, Gavin, 233
 Magali, Richard, 147, 202, 211
 Maigret, Bernard
 Mallet, Vincent, 274

 Malliavin, Therese
 Mamgain, Khushi
 Mandier, Céline, 277
 Mandonnet, Nathalie, 236
 Mantelin, Sophie, 259
 Marbehan, Xavier, **250**
 Marcelino, Isabel, 214
 Marchais, Antonin, 150
 Marchal, Pierre
 Marcovich, Gauthier
 Marechaux, Angélique
 Marenne, Gaëlle, 151, 231
 Mareuil, Fabien, **274**
 Mariadassou, Mahendra, 220, 287
 Marigo, Omar
 Marin, Jean-Michel, 90
 Marin, Romuald, **34**
 Markov, Gabriel, **158**, **249**
 Marku, Malvina, 38
 Marques Da Costa, Maria-Eugenia, 150
 Marsolier, Marie-Claude, 154
 Martin, Luc
 Martin, Michael, 154
 Martin, Véronique, 287
 Martinez Noriega, Mariana
 Martinez-Noriega, Mariana, **222**
 Massau, Karine, 136, **184**
 Mathé-Dehais, Catherine, 192
 Mayeur, Hélène, **177**
 Mazan, Sylvie, 177
 Mazoyer, Clément
 Mbouamboua, Yvon
 Mc Culloch, Elaine, 233
 Mcleer, Anne, 209
 Médigue, Claudine, **227**
 Mehdi, Abdenacer, 284
 Merchadou, Kévin
 Merda, Déborah, **155**, 161, **180**
 Mergez, Alexis, **141**
 Messak, Imane
 Mestivier, Denis
 Meštrović, Nevenka, 149
 Meyer, Damien, **200**, 289
 Meyer, Vincent, 151
 Michel, Léo, 177
 Michel, Mathilde, 257
 Midoux, Cédric, 287
 Mohamed, Anliat
 Moine-Franel, Alexandra, 274
 Molina, Franck, 248
 Moncion, Thomas, 219
 Monget, Philippe, 144
 Moniot, Antoine, 242
 Montout, Laura, 65
 Moquin-Beaudry, Gael, 150
 Morel, V., 151
 Morga, Benjamin, 272
 Morin, Emmanuelle, 178

Morvan, Micks
 Moser, Mirko, 83
 Mottier, Stephanie
 Mourad, Raphaël, 141
 Mouren, Jean-Christophe, 272
 Muller, Jean, 111
 Multari, Maxime
 Mundwiller, Emeline, 197

 Nandkishore, Nitya, 230
 Naouar, Naira, 233
 Naquin, Delphine, 284
 Nasso, Isabelle, 225
 Naves, Michel, **236**, 289
 Nerriere, Virginie
 Neuvéglise, Cécile, 220
 Neves, Aitana, 233
 Nicaise, Samuel, 111
 Nicolaiew, Nathalie
 Noel, Benjamin, 149
 Noël, Cyril, 68, 283
 Nozais, Mathis
 Nuel, Gregory, 72
 Nunes, Flavia, 262

 Ogloblinsky, Marie-Sophie, 208
 Olaso, R., 151
 Oliveira, Luciana, 278
 Olivier, Alibert, 244
 Olmos, Eric, 250
 Orjuela, Julie
 Osipenko, Maria, 237
 Ouazahrou, Rania, **284**
 Ozisik, Ozan, 159, 252

 Paineau, Damien, 291
 Pancaldi, Vera, 38
 Paquis-Fluckinger, Véronique, 203
 Pasquier, Raphaël
 Patrick, Durand, 136
 Paulo De La Reberdiere, Nilda, 289
 Pécréaux, Jacques, 207
 Pelletier, Camille, 272
 Perdry, Hervé, 208
 Péré, Arthur, 149
 Pereira Dos Santos, Mateus, **183**
 Perfus-Barbeoch, Laetitia, 149
 Pericard, Pierre
 Pernet, Alix, 260
 Perrière, Guy
 Peticca, Aurélie, **221**
 Petitjean, Marie
 Pflieger, David
 Pham, Hoang Giang
 Pham, Hoang-Giang, **261**
 Philippe, Cathy, **30**
 Picolo, Floriane, **144**
 Piégu, Benoît, 144
 Pittion, Florence, **147**

 Piucco, Romain, **266**
 Plaza Oñate, Florian, 219
 Plenecassagnes, Julien
 Plocoste, Thomas, 195
 Plomion, Christophe, 182
 Plumain, Didier, 251
 Pochon, Mathis, **259**
 Polit, Lélia
 Pomiès, Lise
 Pons, Marie-Anne, 214
 Popot, Marie-Agnès, 196
 Porracciolo, Paola, 83
 Porro, Barbara, **186**
 Postec, Anaïs
 Pot, Matthieu, 162
 Potier, Delphine, 281
 Pouillet, Nausicaa, **101**
 Poupot, Mary, 38
 Pousse, Mélanie
 Pratella, David, 203
 Prifti, Edi, 206, 219, 285
 Puller, Vadim, 219

 Quesneville, Hadi, 174
 Quétel, Isaure
 Quétel, Isaure, **20**, 162, **214**
 Quinquis, Fabien

 Racoupeau, Martin, **240**
 Rancurel, Corinne, 149
 Raoux, Corentin
 Rastogi, Nalin, 193
 Réda, Clémence
 Redon, Richard, 151
 Reinharz, Vladimir, 36
 Renaud, Yoan
 Renault, Pierre, 220
 Rene-Trouillefou, Malika, 225
 Reveillaud, Julie, 216
 Rioualen, Claire
 Ripoll, Julie, **100**, **277**
 Rique, Flavian, **244**
 Rivals, Eric, **82**, 100, 277
 Rizzon, Carène, 198
 Robbe-Sermesant, Karine, **149**, 259
 Robin, Stéphanie, **270**
 Rocha, Eduardo, 42
 Rogier, Odile
 Rouillon, Marine, **258**
 Rousseau, Baptiste, **282**
 Rousseau, Coralie, **224**
 Rousseau, Jérémy, **235**
 Roussel, Natacha
 Roux, Didier, 214
 Roy, Gaspar, **285**
 Roy, Gaspard
 Rué, Olivier, 220, 287
 Ruminy, Philippe, 191

Sabban, Jules
 Sabot, François, 221
 Sagot, Marie-France
 Saidi, Somia, **174**
 Sailleau, Corinne, 180
 Saint-Hilaire, Maïlie, 251
 Saint-Pierre, Aude, 151
 Salgado, David, 233
 Saliba Albuquerque Freire Érika, Valeria, 183
 Sallaberry, Marine, 259
 Samaran, Flore, 257
 Samson, Franck, **198**
 Samson, Samantha
 Sanchez, Anne-Carmen, 218
 Sanchez-Flores, Fidel-Alejandro, 222
 Sand, Olivier, 281
 Sandron, F., 151
 Santagostini, Pierre, 260
 Sarti, Edoardo
 Sassolas, Fabien, **172**
 Sater, Vincent, 191
 Savescu, Paul
 Sayeb, Maroua, 155
 Schbath, Sophie, 287, **291**
 Scheidecker, Sophie, 111
 Schuler, Hannes, 83
 Scoazec, Jean-Yves, 150
 Sebaoun, Jean-Marc, 151
 Segretier, Wilfried, 193
 Seiler, Julien, 288
 Sémery, Maëla
 Seraphin, Rémi
 Servant, Florence
 Sessegolo, Camille
 Si Ahmed, Yanis, 113
 Sidibé, Ouléye, **218**
 Siegel, Anne, 158, 249
 Simon, Bruno, 232
 Sklab, Youcef, 285
 Smaïl- Tabbone, Malika, 243
 Smith, Caleb
 Soumet, Christophe, 179
 Soun, Camille, **238**
 Spatz, Madeleine, 291
 Sperandio, Olivier, 274
 Stattner, Erick, **193**
 Summo, Marilyne, 200

 Tadrent, Nachida, 134, **226**
 Talarmin, Antoine, 162, 214, 251
 Taly, Antoine
 Tanguy, Gwenn, 224
 Taupin, Jean-Luc, 243
 Tayeh, Nadim, 212
 Téletchéa, Stéphane, 245, 246
 Tellini, Nicolo
 Tenenhaus, Arthur, 146
 Terezol, Morgane, 281
 Térézol, Morgane, **252**

 Terumalai, Dillenn, 233
 Theil, Sébastien, 220
 Thermes, Claude, 284
 Thevenon, Julien, 209
 Thuilliez, Corentin, **150**
 Tibiri, Ezechiël
 Tichit, Laurent
 Tirera, Sourakhata, **25**
 Toffano, Antoine, **207**
 Togawa, Roberto, 183
 Tonazzolli, Arianna, **233**
 Torterotot, Maelle, 257
 Traore, Adriana, 233

 Ule, Jernej, 247
 Umber, Marie, 289
 Uricaru, Raluca
 Urrutia, Ander, **225**
 Usureau, Cédric, 243

 Valdeolivas, Alberto, 159
 Van Blerk, Sebastian
 Van Dijk, Erwin, 284
 Van Ghelder, Cyril, 259
 Van Helden, Jacques, 233, 288
 Vandecasteele, Céline, 259
 Vandel, Jimmy, 281, 290
 Vayssade, Jehan-Antoine
 Vega Rua, Anubis
 Vega-Rua, Anubis, **22**
 Velo Suarez, Lourdes
 Velo-Suarez, Lourdes, 216
 Verstraete, Nina, **38**
 Viailly, Pierre-Julien, 191
 Vialaneix, Nathalie, **31**
 Vidal, Valérie, 287
 Viennot, Mathieu, 191
 Vigo, Elora
 Vila Nova, Meryl, 161
 Vingataramin, Youri, 214
 Volff, Jean-Nicolas, 172

 Wincker, Patrick, 149
 Wörtwein, Tobias, 286

 Xhaard, Constance

 Yao, Jean-Elisée, 203
 Yassine, Mohamad
 Ysebaert, Loïc, 38
 Yvon, Claire, 155

 Zaffran, Stéphane, 230
 Zallio, Mathieu, 186
 Ziane, Khaoula, **229**
 Zientara, Stephan, 180
 Zins, Marie, 151
 Zotta-Mota, Ana-Paula, 149, 176, 183
 Zucker, Jean-Daniel, 206, 285
 Zytnicki, Matthias, 141, 190, 261