UMR IRISA

Activity Report 2025

Team GENSCALE

Scalable, Optimized and Parallel Algorithms for Genomics

Joint team with Centre Inria de l'Université de Rennes

D7 – Data and Knowledge Management

Université de Rennes | rennes | INSA INSTITUT NATIONAL DES SCIENCES APPLIQUÉES RENNES | cnrs | Inria | Inserm | CentraleSupélec | IMT Atlantique Bretagne-Pays de la Loire École Mines-Télécom | Université Bretagne Sud ubs

Project-Team GENSCALE

Creation of the Project-Team: 2025 March 01

Keywords

Computer sciences and digital sciences

A1.1.1. – Multicore, Manycore
A1.1.2. – Hardware accelerators (GPGPU, FPGA, etc.)
A1.1.3. – Memory models
A1.1.4. – High performance computing
A3.1.1. – Modeling, representation
A3.1.2. – Data management, quering and storage
A3.1.8. – Big data (production, storage, transfer)
A3.1.11. – Structured data
A3.3.3. – Big data analysis
A6.3.3. – Data processing
A7.1. – Algorithms
A7.1.2. – Parallel algorithms
A7.1.3. – Graph algorithms
A8.2. – Optimization

Other research topics and application domains

B1.1.4. – Genetics and genomics
B1.1.6. – Evolutionnary biology
B1.1.7. – Bioinformatics
B2.2.4. – Infectious diseases, Virology
B2.2.6. – Neurodegenerative diseases
B2.3. – Epidemiology
B2.4.2. – Drug resistance
B3.5. – Agronomy
B3.6. – Ecology
B3.6.1. – Biodiversity
B9.5.6. – Data science

Contents

Project-Team GENSCALE **1**

1 Team members, visitors, external collaborators **5**

2 Overall objectives **6**

3 Research program **8**
3.1 Axis 1. Make the data exploitable . . . 8
3.1.1 Representation & Compression . . . 9
3.1.2 Indexation . . . 9
3.2 Axis 2. Release the information contained in data . . . 9
3.2.1 Complex variants & pangenome graph analyses . . . 10
3.2.2 Metagenome analyses . . . 10
3.2.3 Recover DNA storage data . . . 11
3.3 Axis 3. Frugal genomic computing . . . 11
3.3.1 Reduce data movements . . . 11
3.3.2 Increase scaling, decrease computations . . . 12

4 Application domains **12**

5 Social and environmental responsibility **12**
5.1 Impact of research results . . . 12
5.2 Footprint of research activities . . . 13

6 Highlights of the year **13**

7 Latest software developments, platforms, open data **13**
7.1 Latest software developments . . . 13
7.1.1 logan-search . . . 13
7.1.2 muset . . . 14
7.1.3 KmerCamel . . . 14
7.1.4 Phylign . . . 14
7.1.5 Phylign-Fulgor . . . 14
7.1.6 MiniPhy . . . 15
7.1.7 FMSI . . . 15
7.1.8 strainberry2 . . . 16
7.1.9 rs-pancat-compare . . . 16
7.1.10 Mapler . . . 16
7.1.11 Alice . . . 16
7.1.12 DnarXiv . . . 17
7.1.13 INVPG_annot . . . 17
7.1.14 SVJedi-Tag . . . 17
7.1.15 ConCluD . . . 18
7.1.16 Kaminari . . . 18
7.1.17 kmindex . . . 18
7.1.18 back to sequences . . . 19

8 New results **19**
8.1 Axis 1. Make the data exploitable . . . 19
8.1.1 Phylogenetic compression . . . 19
8.1.2 Logan Search, a k-mer search engine for all Sequence Read Archive public accessions 19
8.1.3 Kaminari, minimizing genomic index sizes . . . 20
8.1.4 K-mer set representation using masked superstrings . . . 20
8.1.5 Optimized k-mer search across millions of bacterial genomes on laptops . . . 20

8.1.6 k-mer matrix compression . . . 21
8.1.7 Optimized phylogenetic batching of million-genome collections for reduced storage requirements and faster data retrieval . . . 21
8.1.8 Towards space-efficient data structures for large genome-distance matrices with quick retrieval . . . 22
8.2 Axis 2. Release the information contained in data . . . 22
8.2.1 Pairwise graph edit distance characterizes the impact of the construction method on pangenome graphs . . . 22
8.2.2 SVJedi-Tag: a novel method for genotyping large inversions with linked-read data . 22
8.2.3 Mapler: a pipeline for assessing assembly quality in taxonomically rich metagenomes sequenced with HiFi reads . . . 23
8.2.4 Investigating pre-assembly clustering of PacBio HiFi reads for *de novo* assembly of complex metagenomes . . . 23
8.2.5 MUSET: set of utilities for constructing abundance unitig matrices from sequencing data . . . 24
8.2.6 Improved strain-level metagenome assembly for modern long reads . . . 24
8.2.7 Strain-level metagenomic classification through assembly-driven database reduction 24
8.2.8 Sequencing data processing for DNA storage . . . 25
8.2.9 Data encoding for DNA storage . . . 25
8.3 Axis 3. Frugal genomic computing . . . 25
8.3.1 Processing-in-Memory: protein database search . . . 25
8.3.2 Processing-in-Memory: energy efficiency . . . 26
8.4 Benchmarks and Reviews . . . 26
8.4.1 Investigating the topological motifs of inversions in pangenome graphs . . . 26
8.4.2 Detecting chromosomal inversions for population genomics: what could be the optimal approach? . . . 27
8.4.3 A review and roadmap for the adoption of pangenomics in agronomy . . . 27
8.5 Applications and bioinformatics analyses . . . 27
8.5.1 Characterization of the newborn microbiota and prediction of metabolite production 27
8.5.2 Accurate MAG reconstruction from complex soil microbiome through combined short- and HiFi long-reads metagenomics . . . 28
8.5.3 Novel genes arising from genomic deletions across the bacterial tree of life . . . 28
8.5.4 Diversity of genomic structural variation across the Tree of Life . . . 28
8.5.5 The Silene latifolia genome and its giant Y chromosome . . . 29
8.5.6 Genomics and transcriptomics of Brassicaceae plants and agro-ecosystem insects . 29
8.5.7 Prediction of genetic relatedness of Escherichia coli using neighbour typing: A tool for rapid outbreak detection . . . 30
8.5.8 Neighbour Typing Using Long Read Sequencing Provides Rapid Prediction of Sequence Type and Antimicrobial Susceptibility of Klebsiella pneumoniae . . . 30

9 Bilateral contracts and grants with industry 31

10 Partnerships and cooperations 31
10.1 International initiatives . . . 31
10.1.1 Participation in other International Programs . . . 31
10.2 International research visitors . . . 32
10.2.1 Visits of international scientists . . . 32
10.2.2 Visits to international teams . . . 32
10.3 European initiatives . . . 32
10.3.1 Other european programs/initiatives . . . 32
10.4 National initiatives . . . 33
10.4.1 PEPR . . . 33
10.4.2 ANR . . . 35
10.4.3 Inria Exploratory Action . . . 37
10.5 Regional initiatives . . . 37

11 Dissemination **38**
11.1 Promoting scientific activities . . . 38
11.1.1 Scientific events: organisation . . . 38
11.1.2 Scientific events: selection . . . 38
11.1.3 Journal . . . 39
11.1.4 Invited talks . . . 39
11.1.5 Leadership within the scientific community . . . 39
11.1.6 Scientific expertise . . . 39
11.1.7 Research administration . . . 40
11.2 Teaching - Supervision - Juries - Educational and pedagogical outreach . . . 40
11.2.1 Teaching administration . . . 40
11.2.2 Teaching . . . 40
11.2.3 Supervision . . . 41
11.2.4 Juries . . . 41
11.3 Popularization . . . 42
11.3.1 Specific official responsibilities in science outreach structures . . . 42
11.3.2 Productions (articles, videos, podcasts, serious games, ...) . . . 42
11.3.3 Participation in Live events . . . 43

12 Scientific production **43**
12.1 Major publications . . . 43
12.2 Publications of the year . . . 43
12.3 Cited publications . . . 48

1 Team members, visitors, external collaborators

Research Scientists

- Pierre Peterlongo [Team leader, INRIA, Senior Researcher, HDR]
- Karel Břinda [INRIA, ISFP]
- Dominique Lavenier [CNRS, Senior Researcher, HDR]
- Claire Lemaitre [INRIA, Senior Researcher, HDR]
- Jacques Nicolas [INRIA, Senior Researcher, HDR]
- Riccardo Vicedomini [CNRS, Researcher]

Faculty Members

- Roumen Andonov [UNIV RENNES, Professor, Emeritus, HDR]
- Khodor Hannoush [UNIV RENNES, until Aug 2025]

Post-Doctoral Fellow

- Loren Dejoies [INRIA, Post-Doctoral Fellow, until Jul 2025]

PhD Students

- Leo Ackermann [INRIA]
- Lune Angevin [UNIV RENNES]
- Siegfried Dubois [INRAE]
- Victor Levallois [INRIA]
- Nicolas Maurice [INRIA]
- Meven Mognol [UNIV RENNES, CIFRE, until Mar 2025]
- Alix Regnier [INRIA]
- Melody Temperville [UNIV RENNES]
- Khac Minh Tam Truong [UNIV RENNES]

Technical Staff

- Charly Airault [CNRS, Engineer]
- Sebastien Bellenous [INRIA, Engineer, from May 2025]
- Olivier Boulle [CNRS, Engineer]
- Julien Leblanc [CNRS, Engineer]
- Meven Mognol [CNRS, Engineer, from Apr 2025]
- Florestan de Moor [CNRS, Engineer]

Interns and Apprentices

- Camille Bourdois [INRIA, Intern, from Apr 2025 until Jun 2025]
- Marcus Foin [INRIA, Intern, from May 2025 until Jul 2025]
- Abel Fresneau [INRIA, Intern, until Jul 2025]
- Marie Picard [ENS RENNES, Intern, from Oct 2025]
- Marie Picard [ENS Rennes, Intern, from May 2025 until Jul 2025]

Administrative Assistant

- Marie Le Roic [INRIA]

Visiting Scientist

- Josipa Lipovac [University of Zagreb, until Apr 2025]

External Collaborators

- Francesca Brunetti [Sapienza University of Rome, external PhD student]
- Erwan Drezen [INSTITUT PASTEUR, until Mar 2025]
- Fabrice Legeai [INRAE]
- Emeline Roux [UNIV RENNES]

2 Overall objectives

Context

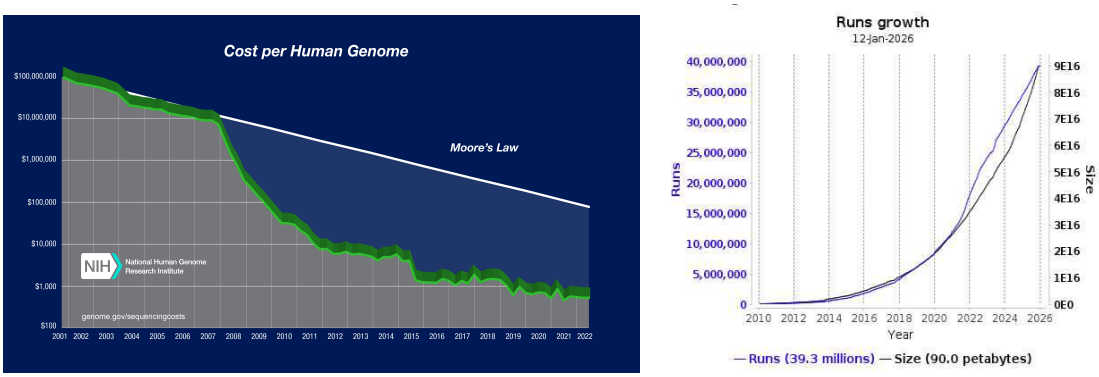
DNA sequencing allows us to read the genetic material of living organisms. In recent decades, there have been multiple successive revolutions in sequencing technologies that have made them widely available. Having access to genomic information has enabled significant progress in fundamental areas, particularly agriculture, environment, and healthcare. Sequence bioinformatics, at the heart of team `GenScale`'s research, is essential to enable and pursue this progress.

The evolution of technologies and practices requires the creation of new algorithmic results to enable the extraction and dissemination of this knowledge. The genomic landscape is evolving. The cost of sequencing has dropped dramatically. For instance, the sequencing price for a human genome is now approximately one thousand euros (see Figure 1 (a)). Consequently, at the same time, the **amount of data** generated and available increases exponentially (Figure 1 (b)). Parallel to the continuous evolution and arrival of novel sequencing techniques, the **data specificity** also evolves. Most recent sequencers produce long sequences (up to millions of characters), and their accuracy increases, up to 99.9% (0.1% of the characters are erroneous), while this error rate was roughly 10% a few years ago.

In addition to the evolution of sequencing techniques, the **scope of research** is rapidly changing. Initially, most studies were tailored to target-sequencing (focusing on a small region of interest of the genome), to a few whole-genomes or to some metagenomes[1]. Today, there are numerous large-scale projects that cover thousands of genomes [71, 78], metagenomes [100], and pangenomes[2] [103, 86]. At the same time, the focus of studies is shifting towards increasingly complex variants.

The current situation is somewhat paradoxical. On the one hand, the diversity, quality and quantity of data is reaching unprecedented levels. On the other hand, biologists do not have the tools they need to exploit this data to the full, due to its size (generic compression methods are not adapted), accessibility (the data is

[1]Metagenome: sequencing of all species of an environmental sample such as soil or seawater
[2]Pangenome: representation of the genome of many individuals in a unique data structure



(a) **(b)**

Figure 1: **(a)** Sequencing cost per human genome. Source: NIH. **(b)** Size of the European Nucleotide Archive databank. Source: EBI.

not indexed globally), and analysis (methods for reconstructing "complex" genomes and metagenomes or detecting large variants are not yet mature).

Our overall aim is to overcome these limitations, by proposing new algorithmic approaches and open-source tools for use by research teams in biology, health, the environment and agronomy. `GenScale` provides the tools to exploit either raw data or massive collections of genomes. These tools enable biologists and bioanalysts to better understand global phenomena such as antibiotic resistance or the adaptation of species to changes in their environment. Similarly, new tools adapted to the constant evolution of sequencing technologies enable us to gain a detailed understanding of variations between genomes, which today are mainly limited to simple point variants representing only 0.1% of a human genome, while structural variants cover 5 to 10 times more base pairs.

An emerging field, data storage on DNA, is facing the same difficulties. It makes intensive use of sequencing technologies to read back the information stored in DNA. The process of sequencing is triggered by each reading of documents, producing a large quantity of sequences that, ideally, should be processed as a stream. The expertise of bioinformaticians is expected to handle these streams of information, even if the 'biology' dimension is not present. The bioinformatics algorithms used to process gigabytes of sequencing data can be advantageously adapted to meet the performance requirements for DNA storage.

The `GenScale` team members have the specificity to cover the whole 'bioinformatics spectrum', from computer-scientists having a strong interest in biological applications to biologists with a deep understanding and taste in algorithmics and their implementations. To our knowledge, this is unique in the international landscape. Our previous results have shown that this configuration offers a fantastic framework for creating impactful tools broadly recognized and utilized in the long term. Our expertise in algorithm and data structure development and their direct application to current genomics problems give the team a strong advantage over international competitors pursuing similar goals.

General objective

The mass and the richness of public data must be a strength, not a threat. The `GenScale` team proposes new high-performance tools to make the most of these masses of genomic data – whether pre-processed or directly from sequencers – for mining them, for detecting subtle signals, and for representing and giving access to the knowledge they hold. Those realizations offer biologists the keys for understanding diseases or resistance traits, at a local scale of one or several species, and at a global scale segregating the information from millions of experiments.

The `GenScale` team collaborates on a daily basis with biologists, some being housed in the team as external collaborators, to validate its methodological contributions and discover new biological information by applying methods it proposes. Real biological data and related questions are fully integrated into the research process. Continuous effort is made to assess the environmental consequences of utilizing our tools. Our objective is to provide tools that are as resource-efficient as possible, suitable for use on basic machines.

Scientific challenges

The research activities are performed through three main research axes and subsequent sub-axes.

- **Axis 1:** Make the data exploitable
  - Representation & Compression
  - Indexation
- **Axis 2:** Release the information contained in data
  - Complex variants & pangenome graph analyses
  - Metagenome analyses
  - Recover DNA storage data
  - Life-scale genomics
- **Axis 3:** Frugal genomic computing
  - Reduce data movements
  - Increase scaling

The full, free and global exploitation of the knowledge contained in genomics data is intimately linked to our ability to easily represent and access these data (axis 1). This objective, also linked with our ability to scale to PB-sized datasets is the purpose of the axis 2.

The scope of genomics projects is changing globally. Although individual genomes remain a primary focus of studies, the trend is towards linking many different entities. Also, the DNA becomes itself a storage support for which dedicated methods have to be developed. All researches conducted in the axis 2 are dedicated to this global trend.

The axis 3 is independent but intersects with all of our research. In view of the sheer volume of data involved and/or the complexity of the problems tackled, it is essential that the software we develop scales terabytes to petabyte sized problem instances if our research is to have any real impact. In addition to algorithmic development, it is clearly essential that the final implementations meet hardware constraints and take full advantage of the functionalities offered by general or specialized hardware.

Additionally, minimizing the impact of our tools on the environment receives special focus. We give special attention to their implementation, optimizing the hardware architectures at our disposal and minimizing the requirement for consumers to purchase new equipment.

Across the board, every research project the `GenScale` team is carried out in collaboration with the biologists community. The presence within the team of INRAE and Numecan collaborators, our close relationships with the national network of environmental, health and agronomical research teams is essential for identifying and developing the right methods with high potential in these fundamental areas of life sciences. The `GenScale` team also benefits from close relations with the `GenOuest` platform. `GenOuest`, initially hosted in the first bioinformatics team at Inria Rennes, offers development, expertise and resources for bioinformatics. It is a major player in the regional bioinformatics community. These collaborations also provide privileged access to data (also enabled by the `GenOuest` platform), offering a playground for developing, testing and validating our approaches, while participating in large-scale applied research.

3 Research program

3.1 Axis 1. Make the data exploitable

The primary goal is to develop novel computational techniques that allow researchers to analyze public sequencing data at the petabyte scale. As sequencing technologies rapidly evolve, the volume of produced data grows at an exponential rate, outpacing the improvements in computational capacities. This growth makes data storage, transmission, and analysis increasingly challenging. For instance, the proportion of sequenced bacteria searchable by the Basic Local Alignment Search Tool (BLAST) [64], the main search

tool of computational biology, is diminishing exponentially, making the search results unrepresentative of the known biological diversity [1]. This lack of searchability is due to current algorithmic and practical infrastructure limitations. It has profound implications across computational biology, medicine, and public health, especially in scenarios requiring rapid decision making, such as during the emergence of a pandemic, in rapid diagnostics of infection diseases, or for the detection of biothreats.

Our research focuses on developing methods to make public genomic data searchable. This includes developing novel space-efficient data representations with varying levels of expressiveness, methods for their compression, and novel techniques for their indexing and querying at a large scale [2]. Our primary target are large public databases, such as SRA [80].

3.1.1 Representation & Compression

Participants: Karel Brinda, Pierre Peterlongo, Alix Regnier, Leo Ackermann, Sebastien Bellenous, Khac Minh Tam Truong.

A challenge is to develop novel data representations for genomic data across the spectrum of used sequencing technologies. This involves implementing a controlled and measured level of lossiness, ranging from lossless capture to very sparse sketches.

Another challenge is to develop scalable methods for high-efficiency compression of genomic data representations. The key idea is to leverage the specific structural properties of genomic data, taking their evolution over time into account. Following the recent development of sequencing technologies, especially the advances in metagenome-assembled genomes, we anticipate that in the next ten years, we will first observe a rapid increase of the total sequenced diversity, followed by a saturation. We aim to develop computationally tractable algorithms to quantify the geometric and repetitiveness measures on public repositories such as SRA in function of time, in order to parameterize the data and exploit their redundancies. For instance, we aim to leverage the phylogenetic relationships between individual constituents of sequencing experiments (genomic, metagenomic, pangenomic, etc.) to guide compression by our technique called phylogenetic compression [1].

3.1.2 Indexation

Participants: Karel Brinda, Pierre Peterlongo, Alix Regnier, Victor Levallois, Sebastien Bellenous.

The knowledge offered by the sequencing data is priceless. Fortunately, a large majority of the sequenced data is freely accessible through large databanks such as ENA and SRA, as previously mentioned. However, the raw sequences stored in these data banks are not indexed and therefore cannot be queried efficiently, apart from direct accession lookups. With a few exceptions, these data sets are never revisited.

Despite some recent research efforts for indexing the raw sequencing data, the most visible one being from Zurich group, Switzerland [79], there exist currently no tool able to provide a search engine able to index hundreds of petabytes of data. State-of-the art in this domain was proposed by the historical `GenScale` team with kmtricks [5] and kmindex [4], used to index dozens of terabytes of data and offering instant queries [2].

In this context, our overall objective is to enable the full and global exploitation of the generated sequences, at the petabyte scale and growing exponentially, allowing users to directly perform queries in raw sequencing data on the fly in order to tap into the largest underexploited resource in life sciences. This is both a medium-term project, for indexing the entire SRA database and also a long-term initiative to adapt to future changes in terms of both size and required functionalities.

3.2 Axis 2. Release the information contained in data

The global aim of this axis is to provide algorithms and software programs to answer specific biological questions. The problem here is no longer to represent and store data in memory, but to explore it and seek out the few pieces of information in this vast quantity of sequences that allow the biologist to answer his

biological question. We focus on some biological issues with sequencing data that are particularly challenging and for which we have strong expertise.

3.2.1 Complex variants & pangenome graph analyses

Participants: Lune Angevin, Siegfried Dubois, Claire Lemaitre, Fabrice Legeai, Melody Temperville, Riccardo Vicedomini.

Phenotype variations, such as species adaptations, resistance to stress in plants, or susceptibility to disease, are directly linked to genomic variations. Therefore, genomic variants have received considerable research efforts for their detection and association with phenotypes. Most of the previous studies focused on small variants (Single Nucleotide Polymorphisms or small insertions and deletions). Recent advances in sequencing technologies have highlighted the prevalence of more complex variations, called structural variants (SVs) that are duplicated, deleted, inverted, or displaced DNA segments, which cover 5 to 10 times more bases than punctual ones [90]. Structural variants have been shown to have a major impact on various phenotypes [104, 63] and on the adaptation and evolution of species [105].

The rise of 3rd-generation sequencing (long reads) has made it possible to characterize and catalog the full range of SVs in many model organisms, such as in humans. However, their genome-wide detection remains difficult when it comes to more complex genomes, such as polyploid ones, more divergent populations, and/or particular types of structural variants such as those with highly repetitive features. Also, long-read technologies are still expensive and inaccessible for re-sequencing projects where a large number of individuals are required. In these cases, short-read technologies are still used either for the quantification problem (genotyping) or in association with long-distance information. We develop methods for improving the detection and quantification of structural variants in such specific contexts and/or data types.

A second area of research concerns genome representation models. The bioinformatics community is currently questioning the dogma of the "reference sequence" to represent the genome of a species, given the large number and diversity of variants detected within a single species [87]. The term '*pangenome*' is used to represent all the genomes and variants that characterize a species. The ideal data structure for this representation is a sequence graph where each node represents part of one or several genome(s), and each genome of the species can be read by traversing a particular path in this graph. Several models and tools have already been proposed by the community [77, 73], but many open problems remain. The construction of these graphs currently relies on heuristics, as does their analysis to characterize genetic variants within a population, with results that vary widely from one method to another. Furthermore, the transposition of classic genomic analysis methods using the linear sequence of the reference assembly to approaches that instead use a reference pangenome in the form of a graph of variations is not immediate. Finally, we work on how to represent, detect and genotype structural variants in such a graph, and in particular when they are close, nested or mixed up with the millions of other small punctual variants [6].

3.2.2 Metagenome analyses

Participants: Roumen Andonov, Lune Angevin, Claire Lemaitre, Nicolas Maurice, Pierre Peterlongo, Emeline Roux, Riccardo Vicedomini.

Metagenomics focuses on the analysis of sequencing data derived from a mixture of microorganisms characterizing an environment of interest [75]. The first step in understanding a microbial environment is characterization of the organisms that are there. This problem requires reconstructing the genomes of the sequenced species (the *metagenome assembly*). We work on this problem at different levels of granularity, and we study the functional aspects of a microbiome (*e.g.*, drug resistance, virulence, pathogenicity).

During metagenome assemblies, strains of the same species (usually sharing high sequence identity) are often "collapsed" into species-level consensus sequences, thus hiding strain variability [102]. Our objective is to provide a strain-aware metagenome assembly brings along additional challenges: (i) the actual number of strains is unknown; (ii) strains of the same species might be highly similar; (iii) the abundance can be so low that it becomes hard to distinguish between rare events and sequencing errors.

We aim to develop novel scalable methods for metagenome assembly (both at species and strain resolution) in the case of high-complexity microbial communities. We develop solutions able to exploit (and possibly combine) the full potential of different sequencing technologies such as short-read, long-read, but also chromosome conformation capture-based technologies (e.g., Hi-C) in order to reconstruct complex metagenomes more accurately at both species and strain level. We also provide formal characterizations of the strain resolution problem based, for instance, on graph theory and combinatorial optimization formulations.

3.2.3 Recover DNA storage data

Participants: Olivier Boulle, Florestan de Moor, Dominique Lavenier, Julien Leblanc.

The process of recovering digital data stored on DNA involves a number of steps, including a sequencing stage which, as with genomics studies, generates a huge volume of data that needs to be processed for downstream analysis [68, 94, 67].

The aim of this axis of research is to propose methods and algorithms that can process the information output while sequencing DNA storing digital data in a time compatible with the overall decoding chain of DNA storage. Unlike biological data, the structure of the information is known. This means that the usual bioinformatics algorithms can be highly optimized to increase their efficiency. We work on this topic in collaboration with researchers who are proposing coding schemes specific to DNA, so that the coding specificities can be used to accelerate the processing of sequencing data. The overall objective is to enable decoding to be carried out on the fly.

3.3 Axis 3. Frugal genomic computing

This research axis aims to limit the environmental usage of the exploitation of bioinformatics tools.

3.3.1 Reduce data movements

Participants: Charly Airault, Karel Brinda, Dominique Lavenier, Meven Mognol, Pierre Peterlongo.

In digital systems, a large part of the power consumption comes from the data movements between the computing units (CPU cores) and the main memory (DIMM modules) [91]. Computing is cheap but transferring data to the computation unit is expensive: reading a 32-bit word from the DRAM, for example, is 10,000 times more energy-intensive than a 32-bit integer addition! And genomic processing often has to navigate randomly through large data structures, leading to low CPU cache efficiency.

One way of mitigating this effect is to design genomic data structures that optimize the locality of data access. A typical example is the Bloom filter, a data structure widely used in genomics to query the membership of an element in a set. Because of their large size, they are generally not cacheable. Knowing whether an element is present requires several accesses (a half-dozen) to an array, which translates into as many cache misses. Optimized Bloom filters, at the cost of a slight degradation in precision, allow an element to target the same memory zone, making much better use of caches and thus limiting data transfer to the CPU [89].

A second approach is the concept of “Processing-in-Memory” (PiM), where the computing units are directly integrated into the main memory [72]. In this case, data no longer leave the memory, but are processed on the spot. In this way, numerous data movements are avoided. The best example of this is probably querying a database stored in a PiM memory: queries are broadcast to all the memory computing units, which process them simultaneously, and send back only relevant data to the CPU [81, 3]. The PiM concept strongly limits the exchanges between the processor cores and the main memory, and contributes to save energy.

To sum-up, this axis has a long term objective of investigating how genomic data structures can be designed to support locality of data access, and exploiting specialized hardware.

3.3.2 Increase scaling, decrease computations

Participants: Pierre Peterlongo, Victor Levallois.

Our research is driven by biological questions involving the analysis of large datasets. The tools we develop, independently from their optimizations, are designed to analyze complete datasets and deliver extensive results. Yet there are many questions that can be answered without the need for complete data analysis, or exhaustive exploration of the associated data structures.

In this context, various methods have already proved successful. This is the case with the sub-sampling of metagenomic data (SimkaMin [65], Minhash [93]) for comparing samples. This is also applied for selecting some specific elements (such as in PebbleScout [97]) from the full datasets when the question is to index and represent samples.

We exploit this concept following various key ideas. 1/ to which extend can we focus only on some elements (sub-sequences, specific portions of graph data-structures, . . .) for answering a biological question? 2/ conversely, determine and exclude large sets of elements that do not carry information for the question asked (ubiquitous sub-sequences, non variables paths on graphs, . . .).

4 Application domains

The `GenScale` team has two distinct application domains: **Computer Science** and **Bio-analyses**.

Computer science. Expected outcomes are new data structures or APIs that will be integrated into other projects internal or external to the `GenScale` team. In the past, in the historical `GenScale` team, this strategy was successful, for example, through the general purpose assembly library GATB [69], data file formats [70], the general purpose minimal perfect hash function bbhash [88], or the library dedicated to linked-read sequencing data LRez [92] to cite a few.

Bio-analyses tools and results. Expected outcomes are end-user tools that are directly applicable to large problems and / or complex problems. This has been the case for most tools developed by members of the historical `GenScale` team, such as DiscoSnp [101], MindTheGap [95], simka [66], SVJedi-graph [96] or kmindex [83]. The expected applications are also new biological results obtained using our algorithmic development or using expertise in major studies [76, 74, 82].

5 Social and environmental responsibility

5.1 Impact of research results

Insect genomics to reduce phytosanitary product usage. Through its long term collaboration with INRAE IGEPP, GenScale is involved in various genomic projects in the field of agricultural research. In particular, we participate in the genome assembly and analyses of some major agricultural pests or their natural enemies such as parasitoids. The long term objective of these genomic studies is to develop control strategies to limit population outbreaks and the dissemination of plant viruses they frequently transmit, while reducing the use of phytosanitary products.

Energy efficient genomic computation through Processing-in-Memory. All current computing platforms are designed following the von Neumann architecture principles, originated in the 1940s, that separate computing units (CPU) from memory and storage. Processing-in-memory (PIM) is expected to fundamentally change the way we design computers in the near future. These technologies consist of processing capability tightly coupled with memory and storage devices. As opposed to bringing all data into a centralized processor, which is far away from the data storage and is bottlenecked by the latency (time to access), the bandwidth (data transfer throughput) to access this storage, and energy required to both transfer

and process the data, in-memory computing technologies enable processing of the data directly where it resides, without requiring movement of the data, thereby greatly improving the performance and energy efficiency of processing of massive amounts of data potentially by orders of magnitude. This technology is currently under test in GenScale with a revolutionary memory component developed by the UPMEM company. Several genomic algorithms have been parallelized on UPMEM systems, and we demonstrated significant energy gains compared to FPGA or GPU accelerators. For comparable performances (in terms of execution time) on large scale genomics applications, UPMEM PIM systems consume 3 to 5 times less energy.

5.2 Footprint of research activities

As our carbon footprint is mainly due to our travels, we propose to monitor the number of train travels, as well as continental and inter-continental flight travels, by `GenScale` members during the current year. Our goal is to measure the evolution of our practices over the years.

Category	**Train travels** (go and return)	**Continental flights** (go and return)	**Intercontinental flights** (go and return)
Permanent researchers (7 persons)	42	6	0
Non-permanent researchers (12 persons)	27	5	0
All researchers (19 persons)	69	11	0

Table 1: Travel statistics in 2025.

6 Highlights of the year

The Nature Methods publication introduced phylogenetic compression – a new principle for representing and searching million-scale microbial genome collections, via using evolutionary history to guide compression and search [8]. The work, originally initiated during the lead author's postdoctoral research at Harvard Medical School and in collaboration with European Bioinformatics Institute, achieves one to two orders of magnitude better compression than generic approaches while remaining fully lossless and compatible with standard k-mer indexes and alignment tools. Today, phylogenetic compression is the core compression technology behind AllTheBacteria [78], the state-of-the-art global collection of 2.4M bacterial genomes, making this massive collection compact, practically distributable, and accessible on ordinary hardware. Since the original paper, multiple contributions from the team have built directly on this foundational work [44, 32, 31, 38, 25], including three PhD theses currently being supervised within the team.

7 Latest software developments, platforms, open data

7.1 Latest software developments

7.1.1 logan-search

Keywords: SRA, Indexing, Genomic sequence

Functional Description: Given a DNA sequence, the service replies in a few minutes in which SRA accession(s) it is likely to occur. The service also enables to recover metadata associated to target accessions, and to perform some visualizations. In more technical depth, the search engine uses kmindex, a k-mer based sequence search tool that uses Bloom filters. It was applied to construct an index over all genome assemblies of all of SRA, more specifically over the unitigs of Logan. This website is running the kmviz visualization tool.

URL: `https://logan-search.org/`

Publication: hal-05446815

Contact: Pierre Peterlongo

Partners: CEA, CHU Pasteur, University of Toronto

7.1.2 muset

Keywords: Unitig matrices, De Bruijn graphs, K-mer

Functional Description: MUSET is a software tool designed to generate an abundance unitig matrix from a collection of input samples in FASTA/Q format. It also offers a comprehensive suite of tools for manipulating k-mer matrices, along with a script for efficiently generating a presence-absence unitig matrix.

URL: https://github.com/camiladuitama/muset

Contact: Riccardo Vicedomini

Participants: Riccardo Vicedomini, Francesco Andreace, Yoann Dufresne, Rayan Chikhi, Camila Duitama

Partner: Sequence Bioinformatics

7.1.3 KmerCamel

Name: KmerCamel

Keywords: Bioinformatics, Compression

Functional Description: KmerCamel provides implementations of several algorithms for efficiently representing a set of k-mers as a masked superstring.

URL: https://github.com/OndrejSladky/kmercamel

Contact: Karel Brinda

7.1.4 Phylign

Name: Phylign

Keywords: Bioinformatics, Alignment, Genomic sequence, Data compression

Functional Description: A tool for rapid BLAST-like search among 661k sequenced bacteria on personal computers.

URL: http://github.com/karel-brinda/mof-search

Contact: Karel Brinda

Participant: Karel Brinda

Partners: European Bioinformatics Institute, HARVARD Medical School

7.1.5 Phylign-Fulgor

Name: Phylign-Fulgor

Keywords: Bioinformatics, Genomics, Alignment, Bacterial strains

Functional Description: Phylign-Fulgor is a Snakemake-based pipeline that enables large-scale local alignment of genomic queries against massive bacterial genome collections such as the 661k and AllTheBacteria datasets on standard laptops and desktops by replacing COBS with a customized version of Fulgor and leveraging phylogenetic compression, which reorganizes genomes according to evolutionary relationships to achieve high compressibility, small index sizes, and efficient k-mer search, it combines fast candidate selection using Fulgor with accurate alignment using Minimap2, supports FASTA/FASTQ inputs, scales through easy parallelization and optional cluster execution, and provides a complete workflow from database download and query preprocessing to matching, candidate aggregation, and final alignments, allowing rapid, memory-efficient, and practical genomic search across hundreds of thousands to millions of bacterial genomes within hours.

URL: https://github.com/Francii-B/Phylign-Fulgor

Contact: Karel Brinda

Participants: Karel Brinda, Francesca Brunetti

7.1.6 MiniPhy

Name: MiniPhy

Keywords: Compression, Bioinformatics, Genomic sequence, Data compression

Functional Description: Phylogenetic compression of extremely large genome collections

URL: https://github.com/karel-brinda/miniphy

Contact: Karel Brinda

7.1.7 FMSI

Name: FMSI

Keywords: Bioinformatics, Genomics, Kmers, Indexing

Functional Description: FMSI is a highly memory-efficient tool for performing membership queries on single k -mer sets. FMSI uses masked superstrings for storing the k -mer sets to ensure high compressibility for a wide range of different k -mer sets, and implements FMS-index, a simplification of the FM-index. It supports both streaming and single queries. The functionality implemented in FMSI is based on the following papers:

The memory consumption for FMSI are (w/o kLCP which is additional 1bit/superstring char):

Queries: 2.41 bits / canonical 31-mer with human genome, $\tilde{3}$ bits / canonical 31-mer for E. coli pangenome (1.17G k-mers from 89k genome), $\tilde{7}$ bits / canonical 31-mer for SARS-CoV-2 Construction: 46 GB for human genome

Release Contributions: FMSI is able to perform dictionary queries, i.e. to return for present k-mers a hash value between 0 and |K|-1 (subcommand hash) FMSI has simplified CLI The parameter for path is now positional Removed some paramaters which were not necessary Query streaming is by default disabled and can be turned on The output of FMSI is now more detailed giving information not only about the percentage of found k-mers, but printing a presence bitmap for each k-mer Due to the changes in CLI and output, this introduces breaking changes compared to v0.3.x

Contact: Karel Brinda

Participant: Karel Brinda

Partner: Charles University Prague

7.1.8 strainberry2

Keywords: Metagenome assembly, Long reads, Haplotype phasing

Functional Description: Strain-level metagenome assembly for long reads obtained with modern sequencing technologies such as PacBio HiFi or Oxford Nanopore R10.

News of the Year: Première version du logiciel.

URL: https://github.com/rvicedomini/strainberry2

Contact: Riccardo Vicedomini

7.1.9 rs-pancat-compare

Keyword: Pangenomics

Functional Description: Program that calculates the distance between two GFA (Graphical Fragment Assembly) files. It takes in the file paths of the two GFA files. The program first identifies the common paths between the two graphs by finding the intersection of their path names. For each common path, the program reads those and output differences in segmentation in-between them. The purpose is to output the necessary operations (merges and splits) required to transform the graph represented by the first GFA file into the graph represented by the second GFA file.

URL: https://github.com/dubssieg/rs-pancat-compare

Publication: hal-04871087

Contact: Siegfried Dubois

Participants: Siegfried Dubois, Claire Lemaitre

Partner: INRAE

7.1.10 Mapler

Name: Metagenome Assembly and Evaluation Pipeline for Long Reads

Keywords: Metagenomics, Genome assembly, Benchmarking, Bioinformatics

Functional Description: Mapler is a pipeline to compare the performances of long-read metagenomic assemblers. The pipeline is focused on assemblers for high fidelity long read sequencing data (e.g. pacBio HiFi), but it supports also assemblers for low-fidelity long reads (ONT, PacBio CLR) and hybrid assemblers. It currently compares metaMDBG, metaflye, Hifiasm-meta, opera-ms and miniasm as assembly tools, and uses reference-based, reference-free and binning-based evalutation metrics. It is implemented in Snakemake.

URL: https://gitlab.inria.fr/mistic/mapler

Publication: hal-04142837

Contact: Nicolas Maurice

Participants: Nicolas Maurice, Claire Lemaitre, Riccardo Vicedomini, Clemence Frioux

7.1.11 Alice

Keywords: Genome assembly, Haplotyping, NGS

Functional Description: Assemble DNA sequencing high-fidelity reads into full genomes. Based on the newly introduced MSR sketching

URL: https://github.com/rolandfaure/alice-asm

Contact: Roland Faure

7.1.12 DnarXiv

Name: dnarXiv project platform

Keywords: Biological sequences, Simulator, Sequence alignment, Error Correction Code

Functional Description: The objective of DnarXiv is to implement a complete system for storing, preserving and retrieving any type of digital document in DNA molecules. The modules include the conversion of the document into DNA sequences, the use of error-correcting codes, the simulation of the synthesis and assembly of DNA fragments, the simulation of the sequencing and basecalling of DNA molecules, and the overall supervision of the system.

URL: https://gitlab.inria.fr/dnarxiv

Contact: Olivier Boulle

Participants: Olivier Boulle, Dominique Lavenier

Partners: IMT Atlantique, Université de Rennes 1

7.1.13 INVPG_annot

Keywords: Pangenomics, Structural Variation, Genomic sequence, Variation graphs

Functional Description: INVPG_annot is an automated method to identify, among the bubbles found in a pangenome graph, the ones that correspond to specific inversion topologies. The method starts with bubbles that have already been found and aims at annotating them, by analyzing their allele size, paths and aligning allele sequences. INVPG-annot takes as input a bubble file in VCF format, and outputs a subset of these bubbles annotated as inversions in VCF format.

URL: https://github.com/SandraLouise/INVPG_annot

Publication: hal-05204329

Contact: Claire Lemaitre

Participants: Claire Lemaitre, Sandra Romain, Siegfried Dubois

7.1.14 SVJedi-Tag

Keywords: Structural Variation, Variation graphs, Genotyping

Functional Description: SVJedi-Tag is a tool for genotyping inversions using linked-read data. It is based on the analysis of the distribution of barcode signals on either sides of inversion breakpoints, with inversions being represented in a variation graph. The variation graph is built from a reference genome and a VCF file containing the inversions to be genotyped. VG giraffe is then used to map the sample reads onto the graph. Then, for each inversion, SVJedi-Tag analyzes the barcode signals of reads aligned on each side of the inversion breakpoints to estimate its allelic ratio and predict its genotype.

URL: https://github.com/Mtemperville/SVJedi-Tag

Publication: hal-05393269

Contact: Melody Temperville

Participants: Fabrice Legeai, Melody Temperville, Claire Lemaitre

7.1.15 ConCluD

Name: ConCluD

Keywords: Bioinformatics, Clustering, Short reads, Long reads, DNA sequencing

Functional Description: DNA-based data storage offers a compelling solution for long-term, high-density archiving. In this framework, accurately reconstructing high-quality encoded sequences after sequencing is critical, as it directly impacts the design of error-correcting codes optimized for DNA storage. Furthermore, efficient and scalable processing is essential to manage the large volumes of data expected in such applications. We introduce a novel method based on de-Bruijn graph partitioning, enabling fast and accurate processing of sequencing data regard less of the underlying sequencing technology and without requiring prior knowledge of the information encoded in the oligonucleotides. It is implemented in C++ within the software ConCluD and optimized for multi-core servers. It also provides Processing-in-Memory acceleration for the UPMEM DIMMs hardware.

URL: https://gitlab.inria.fr/pim/org.pim.dnarxiv

Publication: hal-05375444

Contact: Florestan De Moor

Participants: Dominique Lavenier, Florestan De Moor, Olivier Boulle

7.1.16 Kaminari

Keywords: Genomics, Indexation, Kmer, Algorithm, Search Engine, Computational biology, Biological sequences

Functional Description: Kaminari is a genomic data indexing tool. In other words, it allows you to create a data structure from a set of DNA documents (assembled genomes, sequencing reads, etc.). This data structure can then be queried like a search engine to find the origin of the DNA sequence of interest.

Contact: Victor Levallois

Participants: Victor Levallois, Pierre Peterlongo

7.1.17 kmindex

Keywords: Kmer, Data structures, Indexing

Functional Description: Given a databank D = {S1, ..., Sn}, with each Si being any genomic dataset (genome or raw reads), kmindex allows to compute the percentage of shared k-mers between a query Q and each S in D. It supports multiple datasets and allows searching for each sub-index Di in G = {D1, ..., Dm}. Queries benefit from the findere algorithm. In a few words, findere allows to reduce the false positive rate at query time by querying (s+z)-mers instead of s-mers, which are the indexed words, usually called k-mers. kmindex is a tool for querying sequencing samples indexed using kmtricks.

News of the Year: During the year 2025, the kmindex tool was vastly updated, mainly by Téo Lemane, including the compression methods developed by Alix Regnier.

URL: https://github.com/tlemane/kmindex

Contact: Pierre Peterlongo

Participants: Teo Lemane, Pierre Peterlongo

7.1.18 back to sequences

Keywords: Kmers, Genomic sequence

Functional Description: "back to sequences" is a software program that allows you to find the origin of words of size k (k-mers) in raw sequencing data. This is a common operation when analyzing this type of data.

News of the Year: We updated in 2025 back to sequences. It is now faster, while accepting more input file formats. It's deployment is now much easier.

Contact: Pierre Peterlongo

8 New results

8.1 Axis 1. Make the data exploitable

8.1.1 Phylogenetic compression

Participants: Karel Břinda.

Comprehensive collections approaching millions of sequenced genomes have become central information sources in the life sciences. However, the rapid growth of these collections has made it effectively impossible to search these data using tools such as the Basic Local Alignment Search Tool (BLAST) and its successors. Here, we present a technique called phylogenetic compression, which uses evolutionary history to guide compression and efficiently search large collections of microbial genomes using existing algorithms and data structures. We show that, when applied to modern diverse collections approaching millions of genomes, lossless phylogenetic compression improves the compression ratios of assemblies, de Bruijn graphs and k-mer indexes by one to two orders of magnitude (implemented in MiniPhy 7.1.6). Additionally, we develop a pipeline 7.1.4 for a BLAST-like search over these phylogeny-compressed reference data, and demonstrate it can align genes, plasmids or entire sequencing experiments against all sequenced bacteria until 2019 on ordinary desktop computers within a few hours. Phylogenetic compression has broad applications in computational biology and may provide a fundamental design principle for future genomics infrastructure. [1].

8.1.2 Logan Search, a k-mer search engine for all Sequence Read Archive public accessions

Participants: Pierre Peterlongo.

There are 50 petabases of freely-available DNA sequencing data. We proposed the yet unpublished Logan Search server [46], built using and running thanks to kmtricks [85] and kmindex [83]. This server allows you to search for any DNA sequence in minutes, bringing Earth's largest genomic resource to your fingertips. Under the hood, we built a 1 petabyte k-mer index for all 27 million sequencing datasets in the SRA up to 12-2023. Logan Search transforms any DNA query to its k-mers (with $k = 31$), and it retrieves every dataset containing these k-mers. It's the only service working at this scale. The output datasets are visualized with custom plots in Logan Search, which accesses a harmonized set of query and SRA meta-data including sequencing technology, type of molecule, geographic distribution, and sample origins. Fundamentally, Logan Search returns a list of SRA accessions. To bring closer to the data we also propose a microservice to instantly retrieve contigs matching the search, also visualisable thanks to a blast-like alignment.

8.1.3 Kaminari, minimizing genomic index sizes

Participants: Victor Levallois, Pierre Peterlongo.

The problem of identifying the set of textual documents from a given database containing a query string has been studied in various fields of computing, e.g., in Information Retrieval, Databases, and Computational Biology. With Kaminari, we consider the approximate version of this problem, that is, the result set is allowed to contain some false positive matches (but no false negatives), and focus on the specific case where the indexed documents are DNA strings. We explore an alternative index design based on k-mers minimizers and integer compression methods. We showed in [48] that a careful implementation of this design outperforms previous solutions based on Bloom filters by a wide margin: the index has lower memory footprint and faster query times, while false positive matches have only a minor impact on the ranking of the documents reported. This trend is robust across genomic datasets of different complexity and query workloads.

8.1.4 K-mer set representation using masked superstrings

Participants: Karel Břinda.

The popularity of k-mer-based methods has recently driven the development of compact k-mer-set representations such as simplitigs/Spectrum-Preserving String Sets (SPSS), matchtigs, and eulertigs, which represent k-mer sets via strings that contain individual k-mers as substrings more efficiently than traditional unitigs. Previously, we showed that all these representations correspond to superstrings of the input k-mers and can be generalized within a unified framework called masked superstrings of k-mers, analyzed the computational complexity of this framework, prove NP-hardness for computing both k-mer superstrings and their masks, and design local and global greedy heuristics for practical computation [99] . We implemented these heuristics in the KmerCamel 7.1.3 tool and demonstrate that masked superstrings unify the theory and practice of textual k-mer set representations while providing a flexible basis for application-specific optimization. In 2025, we focused primarily on the optimization of the approximation algorithms, in particular for greedy approximation for simplitigs/Spectrum-Preserving String Sets, which resulted in new major releases of KmerCamel 7.1.3.

Building on this foundation, we addressed the need for scalable and versatile indexes for arbitrary k-mer sets in the face of rapidly growing and heterogeneous genomic data vian an index called FMSI. Unlike state-of-the-art indexes such as SBWT and SSHash, whose performance relies on long non-branching paths in de Bruijn graphs, FMSI adopts a superstring-based design that supports efficient membership and compressed dictionary queries with strong theoretical guarantees. FMSI exploits recent advances in k-mer superstrings together with the Masked Burrows–Wheeler Transform, which extends the classical Burrows–Wheeler Transform by incorporating position masking. Across genomic, pangenomic, and metagenomic datasets and a wide range of k values, FMSI consistently improves query space efficiency, often using 2–3× less memory than competing methods while maintaining comparable or faster query times. [19]

Finally, we further extended the masked superstring framework to support dynamic operations on large k-mer sets, a setting that static compact representations cannot handle efficiently. We introduced f-masked superstrings, which combine masked superstrings with custom demasking functions to enable k-mer set operations through index merging. By integrating this concept withi the FMSI index, we obtained a memory-efficient k-mer membership index and compressed dictionary that support set operations via Burrows–Wheeler Transform merging. This approach offers a promising theoretical solution to dynamic k-mer set manipulation and highlights f-masked superstrings as a strong candidate for an elementary data type for k-mer sets.sladky: [24]

The masked superstring framework was presented RECOMB-2025 [59] and DSB 2025 [36].

8.1.5 Optimized k-mer search across millions of bacterial genomes on laptops

Participants: Karel Břinda, Francesca Brunetti.

Comprehensive bacterial collections have reached millions of genomes, opening new opportunities for point-of-care diagnostics and epidemiological surveillance. However, local real-time search over such collections on commodity hardware remains difficult. Currently, only LexicMap and Phylign enable local search and alignment at such a scale; among them, only Phylign is designed to run on laptops, via a subindex approach informed by phylogenetic compression. However, Phylign's performance deteriorates on long and divergent queries because it uses COBS as a k-mer-based prefilter before alignment with Minimap2. Meanwhile, recent k -mer indexes such as Fulgor and Themisto have emerged, but there is no practical methodology for selecting, combining, and parameterizing them for phylogenetically partitioned million-genome search under constraints. Here, we develop an end-to-end methodology for k-mer matching in phylogenetically compressed bacterial collections. We formalize a matching strategy defined by matching mode, query type, and reference characteristics, and use this to shortlist candidate indexes and benchmark them under space–time trade-offs. As a case study, we address plasmid search over AllTheBacteria, compare multiple index types, and identify configurations optimizing the Pareto frontier of space and speed. Guided by these results, we implement a phylogenetically compressed variant of Fulgor, integrate it into Phylign, and obtain Phylign-Fulgor, a laptop-ready pipeline for million-genome search. On the 661k collection, Phylign-Fulgor makes the prefiltering step 4× faster than Phylign-COBS at the cost of a 1.2× larger index. On AllTheBacteria, its k-mer filter is 20x–300x faster in real time than LexicMap's alignment-based search and uses 20× smaller disk space. The full Phylign-Fulgor workflow including Minimap2 alignments is slower than LexicMap for a single plasmid but competitive or faster for batched plasmid queries. Phylign-Fulgor has comparable matching sensitivity to LexicMap, is less sensitive at the alignment level, but always stays within a laptop RAM budget (5×–20× lower memory than LexicMap). [44] The work was presented at DSB 2025 [31] and 53rd National Congress of the Italian Society of Microbiology [32].

8.1.6 k-mer matrix compression

Participants: Pierre Peterlongo, Alix Regnier.

In the work [35] presented at DSB2025, we propose a block compression method on top of the kmtricks matrices [84], when used by kmindex for indexing and query purposes. The main objective is to achieve sensible compression ratios with a limited impact on the index creation time and on the query time. This method enables partial and targeted decompression, thus optimizing data processing. To improve the compressibility of matrices, we also exploit an idea inspired by phylogenetic compression [8]. Hence, the impact of reordering matrix columns based on phylogenetic order is studied and used as a means of increasing their compressibility. Preliminary results show that this reordering significantly improves matrix compressibility. Finally, the approach we propose reduces the storage space required, with a limited impact on query time, making k-mer matrices more accessible and usable.

8.1.7 Optimized phylogenetic batching of million-genome collections for reduced storage requirements and faster data retrieval

Participants: Tam Truong, Dominique Lavenier, Pierre Peterlongo, Karel Břinda.

Efficient compression is essential for the practical use of million-genome bacterial collections, yet increasing scale and diversity make this progressively harder. The most effective approach to date is phylogenetic compression, which orders genomes according to evolutionary relationships, as implemented in MiniPhy. MiniPhy groups genomes into phylogenetically related batches of limited size and then applies local phylogenetic reordering prior to compression, forming the basis of AllTheBacteria (n = 2,440,377)

and achieving a 40-fold reduction compared to gzip. However, MiniPhy relies on heuristic batching based on accession order within species, used as a proxy for similarity. While sufficient for smaller datasets, this assumption breaks down at million-genome scale, where individual species can exceed hundreds of thousands of genomes and related genomes are no longer close in accession order, leading to reduced compression efficiency. In addition, MiniPhy is tightly coupled to XZ, limiting the use of newer compressors that offer comparable or better compression with substantially faster performance. We introduce a new phylogenetically informed batching strategy based on approximate species-level phylogeny reconstruction, enabling effective co-location of related genomes even at million-genome scale. Applied to AllTheBacteria excluding rare or unknown species (n = 2,282k), this yields batches over 40% more compressible with XZ (20% when all genomes are included). Using MBGC further reduces the compressed size to 15 GB (41.3 GB for the full collection), and combining this with optimized AGC results in an 18× improvement in random-access single-genome retrieval time (0.5 s versus 8.9–9.5 s). [38]

8.1.8 Towards space-efficient data structures for large genome-distance matrices with quick retrieval

Participants: Léo Ackermann, Karel Břinda, Pierre Peterlongo.

Many standard bioinformatics analyses lean on pairwise distances between genomes. As a result, the scalability of multiple downstream analyses relies on the efficient computation and storage of pairwise distance matrices. However, while the efficient computation of distances has been addressed by modern sketching-based methods such as Mash [93] and successors, the storage and indexing of the resulting matrices remain a significant challenge. In fact, due to their quadratic size in the number of genomes, these matrices already surpass most storage capacities and are thus heavily truncated when stored. This calls for a dedicated data structure that would be space-efficient and support near-constant-time distance retrieval queries. In our work, we focused on showed that theoretical collections of genomes model can be stored in linear space supporting constant-time queries. We then demonstrated our preliminary results on the tradeoffs that exist between metric distortion and storing cost in practical use cases. Preliminary results were presented at DSB2025 [25].

8.2 Axis 2. Release the information contained in data

8.2.1 Pairwise graph edit distance characterizes the impact of the construction method on pangenome graphs

Participants: Siegfried Dubois, Claire Lemaitre.

Pangenome variation graphs are an increasingly used tool to perform genome analysis, aiming to replace a linear reference in a wide variety of genomic analyses. The construction of a variation graph from a collection of chromosome-size genome sequences is a difficult task that is generally addressed using a number of heuristics. The question that arises is to what extent the construction method influences the resulting graph, and the characterization of variability. We aimed at characterizing the differences between variation graphs derived from the same set of genomes with a metric which expresses and pinpoint differences. We designed a pairwise variation graph comparison algorithm, which establishes an edit distance between variation graphs, threading the genomes through both graphs. We applied our method to pangenome graphs built from yeast and human chromosome collections, and demonstrated that our method effectively characterizes discordances between pangenome graph construction methods and scales to real datasets [11].

pancat compare is published as free Rust software under the AGPL3.0 open source license. Source code and documentation are publicly available.

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

Participants: Fabrice Legeai, Claire Lemaitre, Mélody temperville.

Structural Variants (SVs) are an important but overlooked aspect of genetic variation. In particular, inversions are known for their role in the evolution of biological diversity and particularly studied in non-model species using population data. One of the major steps in the study of SVs is genotyping. Linkedread data provide a cost-efficient alternative to long-reads to genotype many individuals, by combining the low sequencing cost of short reads with long-distance information thanks to the use of barcodes tagging long molecules. Whereas several methods have been proposed to discover SVs with linked-reads, there are currently no tool for genotyping with this type of sequencing data. In this paper, we present SVJedi-Tag, the first inversion genotyping method dedicated to linked-read data. We tested SVJedi-Tag on simulated and real linked-read data in the seaweed fly *Coelopa frigida*, and showed that SVJedi-Tag is able to genotype with high accuracy large inversions above 25 kb, with a read depth as low as 3X [37].

8.2.3 Mapler: a pipeline for assessing assembly quality in taxonomically rich metagenomes sequenced with HiFi reads

Participants: Claire Lemaitre, Nicolas Maurice, Riccardo Vicedomini.

Metagenome assembly seeks to reconstruct the most high-quality genomes from sequencing data of microbial ecosystems. Despite technological advancements that facilitate assembly, such as Hi-Fi long reads, the process remains challenging in complex environmental samples consisting of hundreds to thousands of populations.

We designed and implemented Mapler [14, 58], a metagenomic assembly and evaluation pipeline with a primary focus on evaluating the quality of HiFi-based metagenome assemblies. It integrates several state-of-the-art tools as well as novel metrics and visualizations based on read-to-contig alignments. It provides a broad view of the sequence characteristics after assembly and binning, in order to identify the bottlenecks faced during bioinformatic processes. It is implemented in an easy-to-use and customizable workflow. We applied Mapler to three metagenomic datasets of increasing complexity, all sequenced using PacBio HiFi technology: a mock community of 21 populations, a human gut microbiome, and a deadwood microbiome, demonstrating that it is an effective way to examine assembly in taxonomically rich ecosystems.

Mapler is open source and publicly available.

8.2.4 Investigating pre-assembly clustering of PacBio HiFi reads for *de novo* assembly of complex metagenomes

Participants: Claire Lemaitre, Nicolas Maurice, Riccardo Vicedomini.

Metagenome assembly remains difficult in taxonomically complex ecosystems, where low-coverage species and graph complexity limit assembly quality. Although clustering reads prior to assembly can reduce graph complexity, it may also lower the effective coverage and has therefore seen limited adoption. This work explored whether upstream clustering of PacBio HiFi long reads can effectively improve metagenomic assembly.

We compared the assembly of individual simulated genomes against those same simulated genomes introduced within a complex metagenome containing related species. We confirmed that genomes are better assembled individually than within the metagenome. Building on this observation, we proposed an iterative clustering-based assembly framework. Reads were first assembled within individual clusters. A complete assembly was than generated through a bottom-up hierarchical approach which aims at recovering mis-clustered or unassembled reads (*e.g.*, reintroducing them into neighboring clusters). The method is modular and compatible with different clustering strategies and assemblers, and can also operate without iteration by simply merging cluster-wise assemblies.

We evaluated this approach on a complex PacBio HiFi soil metagenomic dataset [52, 42] using both coverage-based and taxonomy-driven clustering strategies. While clustering reduced the fraction of reads aligning back to the final assembly and the iterative rescue yielded only marginal improvements, several genomes recovered through clustering-based strategies showed higher quality in terms of completeness, contamination, and contiguity. These results indicate that, despite current limitations in pre-assembly clustering and read rescue, clustering-based approaches could in principle enhance the recovery of specific genomes and warrant further methodological improvements [33].

8.2.5 MUSET: set of utilities for constructing abundance unitig matrices from sequencing data

Participants: Riccardo Vicedomini.

Unitigs are biological sequences that compactly and exhaustively represent sequencing data or assembled genomes. They are constructed from k-mers but they avoid the redundancy of sequences covering the same genomic locus. A unitig matrix is a data structure representing sequence content across multiple experiments by recording a numerical value for each unitig across all samples. Unitig matrices extend the concept of k-mer matrices, which are gaining popularity for sequence-phenotype association studies. They additionally preserve variations between samples while, at the same time, significantly reducing disk space compared to k-mer matrices. We addressed the limitations of current methodologies by developing MUSET [21], a method that integrates efficient k-mer counting and unitig extraction in order to generate unitig matrices containing abundance values across the input samples. MUSET overcomes the limitation of state-of-the-art methods which could only output binary (*i.e.*, presence-absence) matrices. We applied MUSET to a 618-GB collection of ancient oral sequencing samples, for which it is able to build an abundance-based unitig matrix in less than 10 hours and using 20 GB of memory. MUSET is expected to facilitate the extraction of biologically significant sequences, making it a valuable contribution to downstream sequencing data analyses such as genome-wide (or metagenome-wide) association studies.

MUSET is open source and publicly available.

8.2.6 Improved strain-level metagenome assembly for modern long reads

Participants: Riccardo Vicedomini.

Recent long-read sequencing technologies (PacBio HiFi and ONT R10.4) have improved metagenome assembly but still struggle to resolve highly similar bacterial strains (sequence identity greater than 99%). We developed Strainberry2 [39], an improved strain-aware post-processing method that reconstructs strain-specific sequences from a species-level assembly and accurate long reads. Strainberry2 combines read dereplication with an iterative strain-phasing algorithm based on co-occurring single-nucleotide variants. Evaluated on 64 simulated datasets and on the ZymoBIOMICS Gut Microbiome Standard mock community, Strainberry2 consistently reduced strain-level errors (2-7X fewer misassemblies and at least 10X fewer strain-switch errors) compared to state-of-the-art assemblers. With PacBio HiFi data, it matched the accuracy of hifiasm-meta while achieving a more balanced trade-off between contiguity and duplication.

Strainberry2 is open source and publicly available.

8.2.7 Strain-level metagenomic classification through assembly-driven database reduction

Participants: Josipa Lipovac, Riccardo Vicedomini.

Strain-level metagenomic classification is essential for understanding microbial diversity and functional potential, but remains challenging, particularly in the absence of prior knowledge about the composition of the

sample. We developed MADRe [49], a modular and scalable pipeline for long-read strain-level metagenomic classification, enhanced with Metagenome Assembly-Driven Database Reduction. MADRe combines long-read metagenome assembly, contig-to-reference mapping reassignment based on an expectation-maximization algorithm for database reduction, and probabilistic read mapping reassignment to achieve sensitive and precise classification. We extensively evaluated MADRe on simulated datasets, mock communities, and a real anaerobic digester sludge metagenome, demonstrating that it consistently outperforms existing tools by achieving higher precision with reduced false positives. MADRe's design allows users to apply either the database reduction or read classification step individually. Using only the read classification step shows results on par with other tested tools.

8.2.8 Sequencing data processing for DNA storage

Participants: Florestan de Moor, Olivier Boullé, Dominique Lavenier.

DNA-based data storage presents an innovative and effective approach for long-term, high-density archiving. Within this context, the precise reconstruction of encoded sequences after sequencing is of paramount importance, as it directly shapes the development of error-correcting codes tailored for DNA storage. Additionally, the ability to process data efficiently and on a large scale is crucial to handle the vast amounts of information anticipated in these applications.

We have developed a new method based on de Bruijn graph partitioning. This approach enables rapid and accurate processing of sequencing data, regardless of the sequencing technology used, and does not require prior knowledge of the information encoded in the oligonucleotides. When tested on both synthetic and real-world datasets, the method demonstrates outstanding precision and recall [15]

The method is implemented in C++ as part of the ConCluD software, which is optimized for multi-core servers. Our experiments reveal that a dataset comprising 89 million reads, equivalent to a 10 GB FASTA file, can be fully processed in less than one minute on a standard server with 32 cores.

8.2.9 Data encoding for DNA storage

Participants: Dominique Lavenier.

We have proposed a novel encoding method for storing encrypted data in DNA form, addressing key biological constraints such as G-C content, homopolymers, and prohibited nucleotide motifs used for data indexing. Our approach is built on two innovations: first, a dynamic encoding (DE) mechanism that utilizes variable-length DNA codewords to encode binary data while avoiding homopolymers longer than N bases; second, a sliding window strategy that detects prohibited motifs in real time and inserts non-coding bases to prevent their formation. Unlike existing schemes, our method scales effortlessly for high homopolymer thresholds and large sets of prohibited motifs, and it remains independent of the cryptosystem used. We provide a theoretical information rate for our proposal, demonstrating significantly higher performance—particularly in terms of information rate—compared to existing schemes, while maintaining the desired G-C content with extremely high probability. Experimental validation using a DNA storage chain simulator confirms that our method requires a comparable number of sequence copies to other encoding strategies for error-free data recovery, underscoring its robustness and efficiency for secure, long-term DNA-based data storage [7].

8.3 Axis 3. Frugal genomic computing

8.3.1 Processing-in-Memory: protein database search

Participants: Charly Airault, Florestan de Moor, Erwan Drezen, Meven Mognol, Dominique Lavenier.

We investigated a massively parallel approach for accelerating protein database search using a Processing-in-Memory (PiM) architecture, leveraging UPMEM's specialized DRAM technology. The study focuses on parallelizing sequence alignment searches within large protein databases, such as UniProt/SwissProt, by distributing the database across PiM modules and broadcasting queries to each processing unit. The PiM-based implementation achieves significant speedups—up to 47.8x faster than traditional multi-core servers—while maintaining result quality comparable to the widely used BLASTp tool. Energy efficiency is also substantially improved, with an 8-core server equipped with 16 PiM modules consuming 13 times less energy for equivalent workloads. Future work includes extending the method to support gapped alignments and integrating advanced search heuristics. This research underscores the potential of PiM technology to revolutionize computationally intensive genomic data processing [26].

8.3.2 Processing-in-Memory: energy efficiency

Participants: Florestan de Moor, Erwan Drezen, Meven Mognol, Dominique Lavenier.

This work evaluates the energy efficiency of genomics algorithms using Processing-in-Memory (PiM) architectures, which reduce data movement bottlenecks by integrating computation directly within memory. Focusing on the UPMEM PiM platform, we assess the performance of key genomic tasks—sorting, data compression, mapping, sequence comparison, and protein database search—on a server equipped with 2,560 DPUs. Results show that PiM delivers significant energy savings (up to 10.5x) and speedups (up to 15.2x) for applications with fine-grained parallelism and irregular memory access patterns, such as protein database searches. However, tasks like sorting, which benefit from optimized CPU cache and SIMD instructions, see limited gains. Our findings highlight PiM's potential to enhance scalability and sustainability in high-throughput bioinformatics, particularly for data-intensive workflows where traditional architectures struggle with memory bandwidth limitations [34].

8.4 Benchmarks and Reviews

We propose in this section new review and benchmark results. They are mainly tied to the Axis 2 (section 8.2).

8.4.1 Investigating the topological motifs of inversions in pangenome graphs

Participants: Siegfried Dubois, Fabrice Legeai, Claire Lemaitre.

Recent technological advances have accelerated the production of high-quality genome assemblies within species, driving the growing use of pangenome graphs in genetic diversity analyses. These graphs reduce reference bias in read mapping and enhance variant discovery and genotyping from SNPs to Structural Variants (SVs). In pangenome graphs, variants appear as bubbles, which can be detected by dedicated bubble calling tools. Although these tools report essential information on the variant bubbles, such as their position and allele walks in the graph, they do not annotate the type of the detected variants. While simple SNPs, insertions, and deletions are easily distinguishable by allele size, large balanced variants like inversions are harder to differentiate among the large number of unannotated bubbles. In fact, inversions and other types of large variants remain underexplored in pangenome graph benchmarks and analyses.

In this work we focused on inversions, which have been drawing renewed attention in evolutionary genomics studies in the past years, and aimed to assess how this type of variant is handled by state of the art pangenome graphs pipelines. We identified two distinct topological motifs for inversion bubbles: one path-explicit and one alignment-rescued. We developed INVPG_annot an automated method to identify, among the bubbles found in a pangenome graph, the ones that correspond to these specific inversion topologies. We constructed pangenome graphs with both simulated data and real data using four state of the art pipelines,

and assessed the impact of inversion size, punctual genome divergence and haplotype number on inversion representation and accuracy.

Our results reveal substantial differences between pipelines, with many inversions either misrepresented or lost, highlighting major challenges in analyzing inversions in pangenomic approaches [50].

8.4.2 Detecting chromosomal inversions for population genomics: what could be the optimal approach?

Participants: Claire Lemaitre, Mélody Temperville.

Chromosomal inversion has aroused a growing interest as it plays a critical role in evolutionary processes. To understand the genetic architecture of adaptation and the importance of inversions, it is important to characterize the full range of inversions accurately and in a cost-efficient way across individuals and populations. We addressed this objective by comparing the effectiveness of characterizing inversions at population-scale using two sequencing methods: long reads (ONT - PacBio) and linked reads (Haplotagging) in two species, *Coelopa frigida* and *Ciona intestinalis*. Analysing long-read high-coverage data was the most exhaustive approach to detect and genotype inversions of all lengths. It proved to be very efficient to refine the breakpoints of adaptive inversions previously detected by an indirect method. Yet, long-reads cannot scale-up to population-wide studies in a cost-efficient manner. Analysing linked-reads data thus provided a relevant alternative to assess the position and frequency of inversions across many more individuals. Overall, the combination of both datasets provided insights into the distribution of inversions across the genomes of both species and across several populations [30, 29, 53].

8.4.3 A review and roadmap for the adoption of pangenomics in agronomy

Participants: Siegfried Dubois, Fabrice Legeai, Claire Lemaitre.

Pangenomics is transforming the way genomics is done. By using assemblies of multiple complete genomes for the same species, it is possible to depart from a biased, reference-centric, point of view. Represented as graphs, these pangenomes open new paths for genome understanding and improvement of species of agricultural interest. However, transitioning to a graph-based approach is not straightforward, and many tools are still needed for this shift. While numerous papers present both methodological and applied approaches on pangenomics, and many reviews present advantages of these methods, very few resources outline what remains to be done. In this opinion paper, we list the bottlenecks that pangenome users experience once they have built their graphs and the methods that are still required to fully exploit pangenomes, and favor the widespread adoption of this approach[43].

8.5 Applications and bioinformatics analyses

8.5.1 Characterization of the newborn microbiota and prediction of metabolite production

Participants: Lune Angevin, Josipa Lipovac, Pierre Peterlongo, Emeline Roux, Riccardo Vicedomini.

Postnatal brain development is influenced by the gut microbiota and its metabolic activity, which are strongly shaped by early-life diet. When breastfeeding is not possible, improving infant formula through targeted microbial supplementation requires precise identification of bacterial communities and their functional potential, ideally at the strain level. This work focuses on the characterization of the gut microbiota and the prediction of metabolite production in piglets fed with breast milk or infant formula using Nanopore R10.4 sequencing of fecal samples [51, 28]. Long-read taxonomic assignment tools were

evaluated using publicly available real-data mock communities [27]. Since functional diversity often occurs at the strain level, the analysis concentrated on strain-level classifiers such as MADRe [49] and ORI [98]. MADRe is able to identify bacterial strains on very large reference databases but, at the same time, it may produce a non-negligible number of false positives. In contrast, ORI yields fewer false positives at the cost of significantly longer runtimes. To combine their strengths and cope with large reference databases, we designed a pipeline in which MADRe provides a reduced set of candidate genomes and ORI refines the selection. A final filtering step based on gene coverage is then applied to remove remaining false positives. This approach makes it possible to reliably identify strains, at the cost of losing some low-abundance strains. Future work will focus on reconstructing bacterial metabolic networks and predicting the metabolites they may produce.

8.5.2 Accurate MAG reconstruction from complex soil microbiome through combined short- and HiFi long-reads metagenomics

Participants: Claire Lemaitre, Nicolas Maurice, Riccardo Vicedomini.

Advances in high-fidelity long-read (HiFi-LR) sequencing technologies offer unprecedented opportunities to uncover the microbial genomic diversity of complex environments, such as soils. While short-read (SR) sequencing has enabled broad insights at gene-level diversity, the inherently limited read length constrains the reconstruction of complete genomes. Conversely, HiFi-LR sequencing enhances the quality and completeness of metagenome-assembled genomes (MAGs), enabling higher-resolution taxonomic and functional annotation. However, the cost and relatively low throughput of HiFi-LR sequencing can limit genome recovery, particularly at the binning stage, where coverage depth is critical. Here, we present a novel hybrid strategy that differs from classical hybrid assemblies, where SR and LR reads are jointly used at the assembly step. Instead, we use high-depth SR data to improve the binning of HiFi-LR contigs. Using both SR and HiFi-LR metagenomic data generated from a tunnel-cultivated soil sample, we demonstrate that SR-derived coverage information significantly improves the binning of HiFi-LR assemblies. This results in a substantial increase in the number and quality of recovered MAGs compared to using HiFi-LR data alone and an incomparable improvement compared to SR data alone. This approach underscores the potential of leveraging the vast amount of publicly available Illumina metagenomic datasets. Completing existing SR resources with PacBio HiFi sequencing can maximise assembly contiguity and binning accuracy using massive amounts of SR data already generated[42].

8.5.3 Novel genes arising from genomic deletions across the bacterial tree of life

Participants: Karel Břinda.

Bacteria are hosts to enormous genic diversity. How new genes emerge, functionalize, and spread remain longstanding questions. Here, we explore a mechanism by which adaptive deletions fuse distant gene fragments. Unlike other gene birth mechanisms that begin with rare, neutral mutations, these “deletion-born fusions” reach high frequency by hitch-hiking on the deletion. The deletion-driven proliferation of the fusion prolongs the mutational supply within these genes before loss, providing additional opportunities for neofunctionalization. We document one such gene fixing and expressing in a long-term E. coli evolution experiment, and identify additional fusion events in the Mycobacterium tuberculosis-bovis split. Finally, we develop a scalable systematic screen to detect these genes in all 2.4 million public single-isolate genomes and identify deletion-born fusions across the bacterial tree of life. These findings challenge the notion that deletions are solely destructive and highlight their role as potential catalysts for evolutionary innovation. [47].

8.5.4 Diversity of genomic structural variation across the Tree of Life

Participants: Claire Lemaitre.

Understanding why some species have higher levels of genetic diversity than others is central in ecology, and has important implications for the evolution and conservation of species. Current research in evolutionary genomics is largely based on Single Nucleotide Polymorphism (SNP), therefore neglecting the genetic diversity and complexity introduced by Structural Variants (SVs) within populations and across species. Leveraging the Darwin Tree of Life project, which routinely generates PacBio HiFi long reads and high-quality genome assemblies for a broad range of taxa, we aim to thoroughly characterize SV diversity across the tree of life. To achieve this, we developed a Snakemake pipeline for detecting heterozygote SVs, integrating multiple aligners and SV callers to retain only the highest-quality SVs for downstream comparative analysis[55, 56].

Besides the necessary and valuable description of SV distribution and determinants on a large phylogenetic scale, this comprehensive dataset allows us to address key evolutionary questions through the lens of structural variation, such as understanding the variability of (structural) genetic diversity across species and examining the overall role of SVs in evolutionary processes. For example, we investigate how variation in effective population size (Ne) between species drives the distribution of structural genetic diversity at a macro-evolutionary scale. We also test whether and how genetic diversity scales with the distribution of different types of genetic variants including different types of SVs, indels, and SNPs. Genetic diversity also appears to be predicted by the species' biology and ecology. We thus compare species with different life histories (such as longevity and body size), as well as variations in geographic range and population census size, to test whether these factors also apply to SVs. Overall, our integrative analysis, combining genomic and ecological data, should be key to understanding the variability of structural genetic diversity across species and examining the role of SVs in evolutionary processes[54].

8.5.5 The Silene latifolia genome and its giant Y chromosome

Participants: Claire Lemaitre.

In some species, the Y is a tiny chromosome but the dioecious plant *Silene latifolia* has a giant Y chromosome of 550 Mb, which has remained unsequenced so far. We participated in the assembly and comparative analysis of a high-quality male *S. latifolia* genome. Comparative analysis of the sex chromosomes with outgroups showed the Y is surprisingly rearranged and degenerated for a 11 MY-old system. Recombination suppression between X and Y extended in a stepwise process, and triggered a massive accumulation of repeats on the Y, as well as in the non-recombining pericentromeric region of the X, leading to giant sex chromosomes [16].

8.5.6 Genomics and transcriptomics of Brassicaceae plants and agro-ecosystem insects

Participants: Fabrice Legeai.

Through its long-term collaboration with INRAE IGEPP, and its support to the BioInformatics of Agroecosystems Arthropods platform, GenScale is involved in various genomic and transcriptomics projects in the field of agricultural research, and participated in the genome assembly and analyses of some major agricultural pests or their natural enemies. In particular, we perfomed a comparative analysis of olfactory (OR) and gustatory receptor (GR) genes and transposable elements (TE) across 12 aphid genomes with varying host ranges[17]. Our results suggest that TE activity may facilitate functional innovation in GRs while alleviating constraints or pseudogenization in ORs. This study reveals how duplication, selection, and TE dynamics shape gene evolution in insect pests. It also provides the first chromosome-scale genome assembly of *Dysaphis plantaginea*, with comprehensive annotations and functional predictions of OR/GR genes, bridging adaptive evolution with mechanistic insights. We also participated to the annotation of the

large lepidoptera genome of *Hylesia metabus*, and the comparison of its content with 17 additional Saturniidae and Sphingidae genomes suggesting that an accumulation of repetitive elements likely led to the increased size of its genome [18]. Using a transcriptomics approach, we identified genes differentially expressed between gonadal and non-gonadal tissues of *Spodoptera frugiperda*, this study targetted kinetochore genes significantly overexpressed in the gonads than in somatic tissues [13]. Furthermore, we compared the genomes of 46 aphid species, and identified genes evolving faster in species that do not alternate their life cycle (i.e. migrating between highly distinct plant hosts), suggesting that the loss of a complex life cycle leads to reduced selective constraints as a consequence of ecological simplification [20]. We also helped to characterize the presence of *Pseudomonas brassicacearum* harboring a gene able to detoxifying harmful isothiocyanates in various *Brassica napus* genotypes, as well as in the guts of the herbivore *Delia radicum*. As we observed that this bacteria was shared consistently in plant and insect, we demonstrated that plant genotypes can shape the gut microbial communities of its pests by promoting the acquisition of symbiotic bacteria [9]. Finally, we also participated to the production of important crop genomics resources such as three *Brassica oleracea* and two *Brassica rapa* [12], or the assembly of *Dactylis glomerata* and *Medicago sativa*. These two last genomes are particularly difficult to assemble due to their polyploidy (auto tetraploid) and their large sizes (3.2 and 7.5 Gbp respectively) [23].

8.5.7 Prediction of genetic relatedness of Escherichia coli using neighbour typing: A tool for rapid outbreak detection

Participants: Karel Břinda.

Identifying the genetic relatedness of resistant bacterial pathogens in healthcare settings can help identify undetected transmission events and outbreaks. However, current methods are time- and resource-intensive. We evaluated a rapid neighbour typing method paired with long-read sequencing for assessment of genetic relatedness.Utilizing a dataset of primary clinical samples and published isolate data from two outbreaks of Escherichia coli, we applied genomic neighbour typing of long- read sequence data to rapidly estimate genetic relatedness. We assessed the correlation between neighbour typing predicted genetic distance and pairwise genetic distance from short-read draft whole genomes for all sample pairs. Predicted genetic trees using neighbour typing were compared to reference genetic trees generated using mash distances and maximum-likelihood (ML) methods to assess the extent of agreement, along with metrics of cluster similarity (cluster comparability and Baker's gamma index) and tree topology similarity (generalized Robinson-Foulds metric).For all three datasets, we found strong correlations between the reference methods and predicted genetic distances (Spearman's rho=0.75-0.95, $p<0.001$), which improved when using a lineage score informed approach (Spearman's rho=0.93-0.94, $p<0.001$). Predicted genetic trees and clusters from neighbour typing were comparable to those generated using either mashtree or a ML method, with a range of cluster comparability of 85.8%-99.5%, Baker's gamma indices of 0.8-0.95, and generalized Robinson-Foulds values of 0.34-0.8.Pairing the neighbour typing method with long-read sequencing can enable accurate predictions of the relatedness of E. coli samples and isolates, and could potentially be used as a rapid outbreak surveillance tool. [10, 22]

8.5.8 Neighbour Typing Using Long Read Sequencing Provides Rapid Prediction of Sequence Type and Antimicrobial Susceptibility of Klebsiella pneumoniae

Participants: Karel Břinda.

The rapid genome-based diagnostic approach of long read sequencing coupled with neighbor typing offers the potential to improve empiric treatment of infection. However, this approach is still in development, and clinical validation is needed to support its use. In this study, we present an assessment of a neighbour typing method (RASE - resistance associated sequence elements) to predict lineage or sequence type (ST) and antimicrobial susceptibility in real time for Klebsiella pneumoniae sensu lato. We analysed the initial

reads generated during the early phase of long read sequencing from pure culture (n=99), mock communities (n=20) and metagenomic samples (n=20). RASE accurately identified 69.7% and 70% of STs in pure culture and metagenomes, respectively, and identified the STs of the isolates representing the highest proportion in mock communities. Regarding antimicrobial susceptibility prediction, the probability of susceptibility increased to 72% (95% CI 63%-80%) across all tested antibiotics, when RASE predicted susceptibility, and decreased probability of susceptibility to 8.9% (95% CI 6.4%-9.6%) when was indicative of a resistant phenotype. Our study confirmed that genomic neighbor typing in K. pneumoniae sensu lato is capable of providing informative predictions of ST and antibiotic susceptibility in less than ten minutes (after the start of sequencing) with 200-500 reads. IMPORTANCE The growing burden of antimicrobial resistance is leading to high rates of mortality and morbidity worldwide. This situation has made the selection of empirical antibiotic therapy challenging, due to the risk of treatment failure and the overuse of last-resort antibiotics. The development of new sequencing technologies is helping to reduce the waiting time for a microbiological diagnosis, providing information in the early phase of bacterial infections, which could help improve clinical outcomes in a time of rising antimicrobial resistance. In this context, we assessed the performance of RASE (resistance associated sequence elements) in Klebsiella pneumoniae , an opportunistic pathogen frequently associated with nosocomial infections, which can rapidly acquire antibiotic resistance genes. Thus, in our study we provide insights that may aid in the validation of RASE for clinical use [45]. The work was also presented as a poster at Applied Bioinformatics and Public Health Microbiology 2025 [57].

9 Bilateral contracts and grants with industry

Participants: Dominique Lavenier, Meven Mognol.

- UPMEM : The UPMEM company is currently developing new memory devices with embedded computing power (UPMEM web site). GenScale investigates how bioinformatics and genomics algorithms can benefit from these new types of memory. A 3 year PhD CIFRE contract (04/2022-03/2025) has been set up.

10 Partnerships and cooperations

10.1 International initiatives

10.1.1 Participation in other International Programs

Logan

Participants: Pierre Peterlongo.

Title: Planetary-Scale Genome Assembly and Indexation

Partner Institution(s):
- University of Toronto
- The Wellcome Trust Sanger Institute
- University of Washington
- JHU - Johns Hopkins University
- SNU - Seoul National University
- Penn State - Pennsylvania State University
- Memorial Sloane Kettering Cancer Center - New York

Date/Duration: from 2025

Additionnal info/keywords: Informal community working on the exploitation of the SRA genomic data

10.2 International research visitors

10.2.1 Visits of international scientists

Other international visits to the team

Josipa Lipova

Status PhD student

Institution of origin: University of Zagreb

Country: Croatia

Dates: November 2024 – April 2025

Context of the visit: visiting PhD student with Riccardo Vicedomini

Mobility program/type of mobility: University of Zagreb PhD mobility (NextGenerationEU)

10.2.2 Visits to international teams

Research stays abroad

Léo Ackermann

Visited institution: Charles University in Prague

Country: Czech Republic

Dates: Dec 2025, 1.5 weeks

Context of the visit: Visit of the Pavel Veselý group at Charles University

Mobility program/type of mobility: PHC Barrande research stay

Karel Břinda

Visited institution: Charles University in Prague

Country: Czech Republic

Dates: Dec 2025, 1.5 weeks

Context of the visit: Visit of the Pavel Veselý group at Charles University

Mobility program/type of mobility: PHC Barrande research stay

10.3 European initiatives

10.3.1 Other european programs/initiatives

Partnership Hubert Curien (PHC) BARRANDE 2025: EFFIMAS: Efficient Indexing of Large Genome Collections via Masked Superstrings of k-mers

Participants: Léo Ackermann, Karel Břinda, Tam Truong.

Coordinators: Karel Břinda, Pavel Veselý

Inria Contact: Karel Břinda

Duration: From Jan 2025 to Dec 2026

Partners: Computer Science Institute, Faculty of Mathematics and Physics (MFF), Charles University, Prague, Czech Republic

Description: Building on our concept of masked superstrings (Sladky, Vesely, and Brinda 2023, 2024), the goal of this project is to significantly extend our preliminary results and address the main challenges in processing and analyzing large k-mer sets, possibly subjected to sampling or sketching. In particular, we have the following specific research aims: Aim 1: Develop a mathematical measure for quantifying the compactness of k-mer sets, so called „Spectrum-Like Quotient“ (SLQ). Aim 2: Develop a SLQ-based parameterization of k-mer sets and k-mer-based algorithms. Aim 3: Develop techniques for indexing large collections of multiple k-mer sets via masked superstrings. Aim 4: Develop streaming algorithms for efficient comparisons of masked superstring. The results of the collaboration will be presented at international conferences, seminars, and colloquia, as well as at both collaborating institutions. The involved students will participate in international workshops and internships, which will broaden their professional horizons and help them prepare for a career in scientific research.

10.4 National initiatives

10.4.1 PEPR

Project MolecularXiv. Targeted Project 2: From digital data to synthetic DNA

Participants: Olivier Boullé, Dominique Lavenier, Julien Leblanc.

Coordinators: Marc Antonini

Duration: 72 months (from Sept. 2022 to Apr. 2029)

Partners: I3S, LabSTIC, IMT-Atlantique, Irisa/Inria, IPMC, Eurecom

Description: The storage of information on DNA requires to set up complex biotechnological processes that introduce a non-negligible noise during the reading and writing processes. Synthesis, sequencing, storage or manipulation of DNA can introduce errors that can jeopardize the integrity of the stored data. From an information processing point of view, DNA storage can then be seen as a noisy channel for which appropriate codes must be defined. The first challenge of MoleculArXiv-PC2 is to identify coding schemes that efficiently correct the different errors introduced at each biotechnological step under its specific constraints.

A major advantage of storing information on DNA, besides durability, is its very high density, which allows a huge amount of data to be stored in a compact manner. Chunks of data, when stored in the same container, must imperatively be indexed to reconstruct the original information. The same indexes can eventually act as a filter during a selective reading of a subgroup of sequences. Current DNA synthesis technologies produce short fragments of DNA. This strongly limits the useful information that can be carried by each fragment since a significant part of the DNA sequence is reserved for its

identification. A second challenge is to design efficient indexing schemes to allow selective queries on subgroups of data while optimizing the useful information in each fragment.

Third generation sequencing technologies are becoming central in the DNA storage process. They are easy to implement and have the ability to adapt to different polymers. The quality of analysis of the resulting sequencing data will depend on the implementation of new noise models, which will improve the quality of the data coding and decoding. A challenge will be to design algorithms for third generation sequencing data that incorporate known structures of the encoded information.

Project Agroecology and digital technology. Targeted Project: Agrodiv

Participants: Siegfried Dubois, Claire Lemaitre, Pierre Peterlongo, Alix Regnier.

Coordinators: Jérome Salse (INRAE)

Duration: 72 months (from Sept. 2022 to Aug. 2028)

Partners: INRAE Clermont-Ferrand (Jerome Salse), INRAE Toulouse (Matthias Zytnicki), CNRS Grenoble (François Parcy), INRAE Paris-Saclay (Gwendal Restoux) and GenScale Irisa/Inria (Pierre Peterlongo)

Description: To address the constraints of climate change while meeting agroecological objectives, one approach is to efficiently characterize previously untapped genetic diversity stored in ex situ and in situ collections before its utilization in selection. This will be conducted in the AgroDiv project for major animal (rabbits, bees, trout, chickens, pigs, goats, sheep, cattle, etc.) and plant (wheat, corn, sunflower, melon, cabbage, turnip, apricot tree, peas, fava beans, alfalfa, tomatoes, eggplants, apple trees, cherry trees, peach trees, grapevines, etc.) species in French agriculture. The project will thus use and develop cutting-edge genomics and genetics approaches to deeply characterize biological material and evaluate its potential value for future use in the context of agroecological transition and climate change. The Genscale team is involved in two of the six working axes of the project. First, we will aim at developing efficient and user-friendly indexing and search engines to exploit omic data at a broad scale. The key idea is to mine publicly available omic and genomic data, as well as those generated within this project. This encompasses new algorithmic methods and optimized implementations, as well as their large scale application. This work will start early 2024. Secondly, we will develop novel algorithms and tools for characterizing and genotyping structural variations in pangenome graphs built from the genomic resources generated by the project.

Project Agroecology and digital technology. Targeted Project: MISTIC - Computational models of crop plant microbial biodiversity

Participants: Claire Lemaitre, Nicolas Maurice, Riccardo Vicedomini.

Coordinators: David Sherman (Inria, Pléiade)

Duration: 60 months (from Nov. 2022 to Nov. 2027)

Partners: GenScale Irisa/Inria, Inria Pleiade, BioCore Inria-INRAE, INRAE Bordeaux (BioGeco, Biologie du Fruit et Pathologie), INRAE Nice-Institut Sophia Agrobiotech.

Description: MISTIC connects the INRAE's extensive expertise in experimental crop culture systems with Inria's expertise in computation and artificial intelligence, with the goal of developing tools for modeling the microbiomes of crop plants using a systems approach. The microbial communities found on roots and leaves constitute the "dark matter" in the universe of crop plants, hard to observe but absolutely fundamental. The aim of the project is to develop new tools for analyzing multi-omics data, and new spatio-temporal models of microbial communities in crops. GenScale's task is to develop new metagenome assembly tools for these complex communities taking advantages of novel accurate long read technologies.

Project Agroecology and digital technology. Targeted Project: BReIF

Participants: Fabrice Legeai.

Coordinators: Anne-françoise Adam-Blondon (INRAE URGI), Michèle Tixier Boichard (INRAE PSGEN) and Christine Gaspin (INRAE GENOTOUL BIOINFO)

Duration: 60 months (from Jan. 2023 to Dec. 2027)

Partners: AgroBRC-RARe, infrastructure (INRAE, CIRAD, IRD), INRAE Genomique, infrastructure (INRAE), BioinfOmics, infrastructure (INRAE), BioinfOmics, infrastructure (INRAE) and various INRAE, IPGRI, IRD and CIRAD units.

Description: The aim of the project is to build a coherent e-infrastructure supporting data management in line with FAIR and open science principles. It will complete and improve the connection between the data production, management and analysis services of the genomics and bioinformatics platforms and the biological resource centers, all linked to the work environments of the research units. It will ensure the connection with the data management services of the phenotyping infrastructures. GenScale is involved in the integration and representation of "omics" data with graph data structures (WorkPackage 2), as well as in the assembly and annotation of several plant and animal genomes and in the building of pangenome graphs (WorkPackage 3).

10.4.2 ANR

Project SeqDigger: Search engine for genomic sequencing data

Participants: Dominique Lavenier, Claire Lemaitre, Pierre Peterlongo, Lucas Robidou, Riccardo Vicedomini.

Coordinator: Pierre Peterlongo

Duration: 55 months (jan. 2020 – June. 2025)

Partners: Genscale Inria/IRISA Rennes, CEA genoscope, MIO Marseille, Institut Pasteur Paris

Description: The central objective of the SeqDigger project is to provide an ultra fast and user-friendly search engine that compares a query sequence, typically a read or a gene (or a small set of such sequences), against the exhaustive set of all available data corresponding to one or several large-scale metagenomic sequencing project(s), such as New York City metagenome, Human Microbiome Projects

(HMP or MetaHIT), Tara Oceans project, Airborne Environment, etc. This would be the first ever occurrence of such a comprehensive tool, and would strongly benefit the scientific community, from environmental genomics to biomedicine.

website: www.cesgo.org/seqdigger

Project GenoPIM: Processing-in-Memory for Genomics

Participants: Charly Airault, Florestan De Moor, Dominique Lavenier, Meven Mognol.

Coordinator: Dominique Lavenier

Duration: 48 months (Jan. 2022 - Dec. 2025)

Partners: GenScale Inria/Irisa, Pasteur Institute, UPMEM company, Bilkent University

Description: Today, high-throughput DNA sequencing is the main source of data for most genomic applications. Genome sequencing has become part of everyday life to identify, for example, genetic mutations to diagnose rare diseases, or to determine cancer subtypes for guiding treatment options. Currently, genomic data is processed in energy-intensive bioinformatics centers, which must transfer data via Internet, consuming considerable amounts of energy and wasting time. There is therefore a need for fast, energy-efficient and cost-effective technologies to significantly reduce costs, computation time and energy consumption. The GenoPIM project aims to leverage emerging in-memory processing technologies to enable powerful edge computing. The project focuses on co-designing algorithms and data structures commonly used in genomics with PIM to achieve the best cost, energy, and time benefits.

website: genopim.irisa.fr

Project REALL: Real-time read alignment to all bacterial genomes on laptops

Participants: Léo Ackermann, Karel Břinda, Pierre Peterlongo, Khac Minh Tam Truong.

Coordinator: Karel Brinda

Duration: 48 months (Oct. 2024 - Sep. 2028)

Description: Rapid search of DNA sequence data is crucial for our ability to control the spread of infectious diseases. However, this presents a major data science challenge: the exponentially growing sequencing data outpace the development of computational capacities, and the increasing data heterogeneity biases search. The central objective of this project is to pioneer innovative methods for rapid, unbiased search across all bacterial genomes on portable devices, with the ultimate goal of achieving real-time alignment of nanopore reads to all sequenced bacteria on standard laptops during sequencing. This will be achieved through advances in phylogenetic compression and entropy-scaling algorithms, and by a novel technology-agnostic graph genome representation. The developed technology will be deployable worldwide, suitable for settings ranging from research laboratories to points of care, and may greatly accelerate downstream applications such as diagnostics of antibiotic resistance.

10.4.3 Inria Exploratory Action

Défi Inria OmicFinder

Participants: Sebastien Bellenous, Victor Levallois, Pierre Peterlongo, Alix Regnier.

Coordinator: Pierre Peterlongo

Duration: 48 months (May 2023 - May 2027)

Partners: Inria teams: Dyliss, Zenith, Taran.

External partners are CEA-GenoScope, Elixir, Pasteur Institute, Inria Challenge OceanIA, CEA-CNRGH, and Mediterranean Institute of Oceanography.

Description: Genomic data enable critical advances in medicine, ecology, ocean monitoring, and agronomy. Precious sequencing data accumulate exponentially in public genomic data banks such as the ENA. A major limitation is that it is impossible to query these entire data (petabytes of sequences).

In this context, the project aims to provide a novel global search engine making it possible to query nucleotidic sequences against the vast amount of publicly available genomic data. The central algorithmic idea of a genomic search engine is to index and query small exact words (hundreds of billions over millions of datasets), as well as the associated metadata.

website: project.inria.fr/omicfinder

Inria AEx BARD(e): Bacterial Antibiotic Resistance Diagnostics(Enhanced)

Participants: Karel Břinda, Loren Dejoies, Jacques Nicolas.

Coordinator: Karel Brinda

Duration: 36 months (2023-2026)

Description: The objective of this AEx is to explore the computational challenges of resistance diagnostics, using a recently developed technique based on ultra-fast nearest neighbor identification among genomes characterized previously. Challenges include the integration of large and heterogeneous genomic and clinical reference data, the deployment of scalable genomic indexes, as well as the deconvolution of signals of individual bacterial species in real clinical samples.

10.5 Regional initiatives

Rennes Metropole: Allocation d'installation scientifique

Participants: Karel Břinda.

Coordinator: Karel Brinda

Duration: 20 months (2024-2026)

Description: Fast DNA sequence data search is crucial for our ability to control the spread of infectious diseases. However, it represents a significant challenge in data science: exponentially growing sequencing data exceeds the development of computing capabilities. The central goal of this project is to develop sublinear algorithms for searching within collections of bacterial genomes, with the ultimate aim of enabling searches on all sequenced bacteria on standard desktop and laptop computers. This will be achieved through advancements in phylogenetic compression and entropy scale algorithms.

11 Dissemination

11.1 Promoting scientific activities

Participants: Karel Břinda, Siegfried Dubois, Dominique Lavenier, Fabrice Legeai, Claire Lemaitre, Jacques Nicolas, Pierre Peterlongo, Emeline Roux, Riccardo Vicedomini.

11.1.1 Scientific events: organisation

General chair, scientific chair

- journées du PEPI IBIS: national meeting of the INRAE network of bioinformatics, Rennes, Oct 2025 (60 participants, 2 days) [Fabrice Legeai]

Member of the organizing committees

- JOBIM mini-symposium on Pangenomics: "Methods for Interfacing with Graphs of Genomic Sequences (MIGGS): Novel Pangenome Paradigms" (125 participants), July 10 2025, Bordeaux, France [Claire Lemaitre]

- Edible Soft Matter 2025: national meeting of the food physico-chemistry ESM, Rennes, July 2025 (140 participants, 4 days) [Emeline Roux]

11.1.2 Scientific events: selection

Chair of conference program committees

- seqBIM 2025: national meeting of the sequence algorithms GT seqBIM, Nantes, Nov 2025 (60 participants, 2 days) [40] [Claire Lemaitre]

Member of the conference program committees

- seqBIM 2025: national meeting of the sequence algorithms GT seqBIM, Nantes, Nov 2025 [Karel Břinda, Claire Lemaitre, Riccardo Vicedomini]

- Jobim 2025: French symposium of Bioinformatics, July 2025, Bordeaux, France [Claire Lemaitre]

- WBCB 2025: Workshop on Bioinformatics and Computational Biology, Telgárt, Slovakia [Riccardo Vicedomini]

- ISMB/ECCB 2025: Intelligent Systems for Molecular Biology / European Conference on Computational Biology, Liverpool, United Kingdom [Karel Břinda, Riccardo Vicedomini]

Reviewer

- Jobim 2025 [Siegfried Dubois]
- ISMB/ECCB 2025: Intelligent Systems for Molecular Biology / European Conference on Computational Biology, Liverpool, United Kingdom [Pierre Peterlongo]

11.1.3 Journal

Reviewer - reviewing activities

- Bioinformatics [Dominique Lavenier]
- Cell Genomics [Claire Lemaitre]
- Nature Communications [Claire Lemaitre]
- PCI Mathematical and Computational Biology [Karel Břinda]
- GigaSciences [Pierre Peterlongo]

11.1.4 Invited talks

- "From bacteria to bits and back again: faster, more accurate, and smarter diagnostics of antibiotic resistance", Czech-French Science Meetup 2025 at the Czech Embassy in Paris, Oct. 2025 [Karel Břinda]
- "Moving from reference genomes to pangenome graphs: benefits and challenges for the analysis of Structural Variation", BiGre Days: Computational biology seminars in Grenoble, Grenoble, Feb. 2025 [Claire Lemaitre]
- "Archivage de données numériques sur ADN", Séminaire des Archives Départementales d'Ille et Vilaine, Rennes, Dec. 2025 [Dominique Lavenier]
- "Open challenges in pangenomics", at "Biology In Silico" Meetings of IGDR Rennes, Apr. 2025. [Siegfried Dubois]
- "Moving from reference genomes to pangenome graphs: benefits and challenges for the analysis of Structural Variation", at "Biology In Silico" Meetings of IGDR Rennes, Apr. 2025. [Claire Lemaitre]
- "Generative AI: worth the cost?", AI4Dev, Grenoble, Feb. 2025. [Siegfried Dubois]
- "Toward sublimera algorithms for searching large genome databases", at Pavel Veselý's group at Charles University, Czech Republic, Dec. 2025 [Leo Ackermann]

11.1.5 Leadership within the scientific community

- Member of the Scientific Advisory Board of the GDR BIMMM (National Research Group in Molecular Bioinformatics) [Claire Lemaitre]
- Animator of the Sequence Algorithms axis (seqBIM GT) of the GDRs BIMMMM and IFM (National Research Groups in Molecular Bioinformatics and Informatics and Mathematics respectively) (350 French participants) [Claire Lemaitre]
- Member of the PEPR MoleculArxiv Executive Committee [Dominique Lavenier]

11.1.6 Scientific expertise

- Expert at DAEI (Pole expertise international de la Délégation au Affaire Européennes et Internationales), MESR [Dominique Lavenier]

11.1.7 Research administration

- Corresponding member of COERLE (Inria Operational Committee for the assessment of Legal and Ethical risks) [Jacques Nicolas]
- Recruitment committees: member of a Univ Rennes CPJ selection committee [Claire Lemaitre]
- Scientific committees: member of the scientific committee of the INRAE's Plant Health and Environment department [Fabrice Legeai]
- Strategic data referent for the INRAE's Plant Health and Environment department [Fabrice Legeai]

11.2 Teaching - Supervision - Juries - Educational and pedagogical outreach

Participants: Lune Angevin, Léo Ackermann, Roumen Andonov, Karel Břinda, Siegfried Dubois, Dominique Lavenier, Fabrice Legeai, Claire Lemaitre, Pierre Peterlongo, Alix Regnier, Emeline Roux, Melody Temperville, Riccardo Vicedomini.

11.2.1 Teaching administration

- Head of the master's degree "Nutrition Sciences des Aliments" (NSA) of University of Rennes (75 students) [Emeline Roux]

11.2.2 Teaching

- Licence: Linear Programming, 30h, L3 Miage, Univ. Rennes, France [Roumen Andonov]
- Licence : Biochemistry, 90h, L1 and L3, Univ. Rennes, France [Emeline Roux]
- Licence: Object Oriented Programming, 38h, L2, INSA Rennes, France [Nicolas Maurice]
- Licence: Introduction to Data Analysis, 28h, L2 Informatics, Univ. Rennes, France [Siegfried Dubois]
- Licence: Object-Oriented Programming, 28h, L2, Univ. Rennes, France [Alix Regnier]
- Master: Short Introduction to R,22h30, M1 Bioinformatics, Univ. Rennes, France [Mélody Temperville]
- Master: Data Visualisation,9h, M1 Bioinformatics, Univ. Rennes, France [Melody Temperville]
- Master: General Statistics, 33h, M1 Nutrition and Food Science, Univ. Rennes, France [Lune Angevin]
- Master: Experimental Bioinformactics, 24h, M1, ENS Rennes, France [Léo Ackermann, Karel Břinda]
- Master: Sequence Bioinformatics, 42h, M1 Informatics, Univ. Rennes, France [Claire Lemaitre, Pierre Peterlongo, Alix Regnier]
- Master: Algorithms on Sequences, 52h, M2 Bioinformatics, Univ. Rennes, France [Léo Ackermann, Claire Lemaitre, Pierre Peterlongo]
- Master : Biochemistry and microbiology, 130h, M1 and M2, Univ. Rennes, France [Emeline Roux]
- Master : Bioanalysis, 6h, M2 Institut Agro Rennes Angers, France [Fabrice Legeai]

11.2.3 Supervision

- PhD: Meven Mognol, Processing-in-Memory [41], CIFRE, defense: July 2025, [Dominique Lavenier]
- PhD in progress: Siegfried Dubois, Characterizing structural variation in pangenome graphs, Inria (PEPR-ANR Agrodiv), since 15/09/2023, [Claire Lemaitre, T. Faraut, M. Zynticki.]
- PhD in progress: Nicolas Maurice, Sequence algorithmics for de novo genome assembly from complex metagenomic data, Inria (PEPR-ANR Mistic), since 01/10/2023, [Claire Lemaitre, C. Frioux, Riccardo Vicedomini].
- PhD in progress: V. Levallois, Indexing genomic data, Défi Inria OmicFinder, since 01/10/2023, [Pierre Peterlongo].
- PhD in progress: Y. Tirlet (IRISA Dyliss team), Univ Rennes (ANR), Integrative method for multi-omics data analysis with application to the activation and regulation of an endogeneized viral genome in a parasitoid wasp, since 01/05/2023, [E. Becker, O. Dameron, Fabrice Legeai.]
- PhD in progress: Francesca Brunetti (Sapienza University of Rome), Assessment of genetic determinants in Escherichia coli uropathogenic lifestyle and intracellular persistence via optimized k-mer matching of million-genome collections on laptops, since 01/11/2022, [Karel Břinda, Maria Pia Conte]
- PhD in progress: Alix Regnier, Limiting the size of data structures indexing genomic sequences, Inria (PEPR-ANR Agrodiv), since 01/10/2024, [Pierre Peterlongo].
- PhD in progress: Léo Ackermann, Developing efficient algorithms for sublinear search in large genome databases, Inria, since 01/10/2024, [Karel Břinda, Pierre Peterlongo].
- PhD in progress: T. Truong, Computational methods for phylogenetic compression, Univ Rennes, since 01/11/2024, [Karel Břinda, Pierre Peterlongo, Dominique Lavenier].
- PhD in progress: Lune Angevin, Fine characterization of the intestinal microbiota and prediction of metabolite production, Univ Rennes, since 01/10/2024, [Emeline Roux, Riccardo Vicedomini, Pierre Peterlongo].
- PhD in progress: M. Temperville, Methods for characterizing structural variations in genomes with linked-read data, Univ Rennes, since since 01/10/2024, [Fabrice Legeai, Claire Lemaitre, Claire Mérot].

11.2.4 Juries

- *Reviewer of HDR thesis*
  - Antoine Limasset, Univ. Lille, Sept. 2025 [Claire Lemaitre]
  - Raja Appuswamy, Sorbone University, Oct. 2025 [Dominique Lavenier]
- *Member of HDR thesis jury*
  - Raluca Uricaru, Bordeaux University, Fev. 2025 [Pierre Peterlongo]
- *President of PhD thesis jury*
  - Francesco Andreace, Pasteur institute Paris, Jan. 2025, [Pierre Peterlongo]
- *Reviewer of PhD thesis*
  - Rick Wertenbroek, Univ. Lausanne, May 2025 [Dominique Lavenier]
  - Thomas Baudeau, Univ. Lille, July 2025 [Dominique Lavenier]
  - Lea Vandamme, Univ. Lille, Dec. 2025 [Dominique Lavenier]
  - Lucas Parmigiani, Bielefeld University, Germany, Nov 2025 [Pierre Peterlongo]

- *Member of PhD thesis jury*
  - Mael Lefeuvre, Natural history museum Paris, Dec 2025 [Pierre Peterlongo]
  - Nastasija Mijovic, Univ. Montpellier, Dec 2025 [Pierre Peterlongo]
- *Member of PhD thesis committee*
  - Léa Nicolas, Univ Rennes [Claire Lemaitre]
  - Soraya Belbati, Univ Rennes [Claire Lemaitre]
  - Dimple Adiwal, Univ Rennes [Claire Lemaitre]
  - Jules Burgat, Univ Rennes [Claire Lemaitre]
  - Florent Couturier, Univ Bordeaux [Claire Lemaitre]
  - O. Weppe, Univ. Rennes [P. Peterlongo]
  - Silpadas Nedoolil, Univ. Rennes [P. Peterlongo]
  - Mael Coatanhay, Univ. Rennes [P. Peterlongo]
  - H. Gasnier, IMT-A, Univ. Brest [Dominique Lavenier]
  - F. Santoro, UBS, Lorient [Dominique Lavenier]
  - M. Caneve, IMT-A, Univ. Brest [Dominique Lavenier]
  - Raneem Jaafar, Univ. Rennes [Riccardo Vicedomini]
  - Fantine Benoit, Univ. Rennes [Riccardo Vicedomini]
  - Emile Breton, Univ. Rennes [Karel Břinda]
  - Alexandra Jalaber Dupont de Dinechin, Univ. Paris Saclay [Fabrice Legeai]

11.3 Popularization

Participants: Lune Angevin, Karel Břinda, Siegfried Dubois, Dominique Lavenier, Julien Leblanc, Pierre Peterlongo, Emeline Roux.

11.3.1 Specific official responsibilities in science outreach structures

- Member of the Interstice editorial board [Pierre Peterlongo]

11.3.2 Productions (articles, videos, podcasts, serious games, ...)

- Media coverage: Mieux caractériser l'antibiorésistance. Sciences Ouest (2025/9, N○431), a popularization article about our research on antibiotic resistance [Karel Břinda]
- Article: Diagnostiquer plus vite la résistance aux antibiotiques, Inria Emergences, a popularization article about our research on antibiotic resistance [Karel Břinda] [web link]
- Video: DNAmaker: Automation of the design of long molecules [Julien Leblanc]. [youtube link]
- Polularization book: Stocker nos données sur ADN [Dominique Lavenier] [60] [61]
- Thès'en Images Presentation of PhD and a thesis project to a group of young people who had dropped out of school at AFPA, which enabled them to make a video with the association Thès'en Images. One intervention in 2025. [Lune Angevin]
- Article: The conversation Biodiversité alimentaire, microbiote et bien-être : la recherche explore les liens potentiels. [Emeline Roux] [62]
- Game: Unplugged activity: Form'IAdable (projet TIARe), a game about supervised learning and its biases [Leo Ackermann] [web link]

11.3.3 Participation in Live events

- Scientific animator at la fête de la Science aux Champs Libres, Oct. 2025, Rennes [Siegfried Dubois]
- Sciences à la Une Creation of a research project for a sixth grade classroom leading to the production of a scientific podcast. Several interventions in 2025. [Lune Angevin]
- Rendez-vous des Jeunes Mathématiciennes et Informaticiennes Discussion with female high school students about university studies, computer science and mathematics. One intervention in 2025. [Lune Angevin]

12 Scientific production

12.1 Major publications

[1] K. Břinda, L. Lima, S. Pignotti, N. Quinones-Olvera, K. Salikhov, R. Chikhi, G. Kucherov, Z. Iqbal and M. Baym. 'Efficient and robust search of microbial genomes via phylogenetic compression'. In: *Nature Methods* 22 (Apr. 2025), pp. 692–697. DOI: 10.1038/s41592-025-02625-2. URL: https://hal.science/hal-04287842 (cit. on pp. 9, 19).

[2] R. Chikhi, T. Lemane, R. Loll-Krippleber, M. Montoliu-Nerin, B. Raffestin, A. P. Camargo, C. J. Miller, M. B. Fiamenghi, D. P. Agustinho, S. Majidian, G. Autric, M. Hugues, J. Lee, R. Faure, K. D. Curry, J. A. Moura de Sousa, E. P. C. Rocha, D. Koslicki, P. Medvedev, P. Gupta, J. Shen, A. Morales-Tapia, K. Sihuta, P. J. Roy, G. W. Brown, R. C. Edgar, A. Korobeynikov, M. Steinegger, C. A. Lareau, P. Peterlongo and A. Babaian. *Logan: Planetary-Scale Genome Assembly Surveys Life's Diversity*. 31st July 2024. DOI: 10.1101/2024.07.30.605881. URL: https://inria.hal.science/hal-05446815 (cit. on p. 9).

[3] D. Lavenier, R. Cimadomo and R. Jodin. 'Variant Calling Parallelization on Processor-in-Memory Architecture'. In: BIBM 2020 - IEEE International Conference on Bioinformatics and Biomedicine. Virtual, South Korea: IEEE, 16th Dec. 2020, pp. 1–4. URL: https://hal.science/hal-03006764 (cit. on p. 11).

[4] T. Lemane, N. Lezzoche, J. Lecubin, E. Pelletier, M. Lescot, R. Chikhi and P. Peterlongo. 'Indexing and real-time user-friendly queries in terabyte-sized complex genomic datasets with kmindex and ORA'. In: *Nature Computational Science* 4.2 (26th Feb. 2024), pp. 104–109. DOI: 10.1038/s43588-024-00596-6. URL: https://hal.science/hal-04489740 (cit. on p. 9).

[5] T. Lemane, P. Medvedev, R. Chikhi and P. Peterlongo. 'kmtricks: Efficient construction of Bloom filters for large sequencing data collections'. In: *Bioinformatics Advances* (29th Apr. 2022). DOI: 10.1093/bioadv/vbac029. URL: https://inria.hal.science/hal-03166007 (cit. on p. 9).

[6] S. Romain and C. Lemaitre. 'SVJedi-graph: improving the genotyping of close and overlapping structural variants with long reads using a variation graph'. In: *Bioinformatics* 39.Supplement_1 (30th June 2023), pp. 270–278. DOI: 10.1093/bioinformatics/btad237. URL: https://inria.hal.science/hal-04155714 (cit. on p. 10).

12.2 Publications of the year

International journals

[7] C. Berton, G. Coatrieux, H. Gasnier, S. Haddad, R. Bellafqira and D. Lavenier. 'A Dynamic Sliding Window Encoding for Secured DNA Data Storage Compliant With Biological and Indexing Constraints'. In: *IEEE Access* 13 (2025), pp. 66009–66030. DOI: 10.1109/ACCESS.2025.3560387. URL: https://hal.science/hal-05067257 (cit. on p. 25).

[8] K. Břinda, L. Lima, S. Pignotti, N. Quinones-Olvera, K. Salikhov, R. Chikhi, G. Kucherov, Z. Iqbal and M. Baym. 'Efficient and robust search of microbial genomes via phylogenetic compression'. In: *Nature Methods* 22 (Apr. 2025), pp. 692–697. DOI: 10.1038/s41592-025-02625-2. URL: https://hal.science/hal-04287842 (cit. on pp. 13, 21).

[9] J. M. Carpentier, S. Derocles, S. Chéreau, B. Marquer, J. Linglin, L. Lebreton, F. Legeai, N. Vannier, A. M. Cortesero and C. Mougel. 'Contrasting glucosinolate profiles in rapeseed genotypes shape the rhizosphere-insect continuum and microbial detoxification potential in a root herbivore'. In: *Msystems* 10.12 (2025), e01269–25. DOI: 10.1128/msystems.01269-25. URL: https://hal.science/hal-05391655 (cit. on p. 30).

[10] A. C. Carroll, L. Mortimer, H. Ghosh, S. Reuter, H. Grundmannd, K. Břinda, W. P. Hanage, A. Li, A. Paterson, A. Purssell, A. M. Rooney, N. R. Yee, B. Coburn, S. Able-Thomas, M. Antonio, A. Mcgeerg and D. R. Macfadden. 'Prediction of genetic relatedness of Escherichia coli using neighbour typing: A tool for rapid outbreak detection'. In: *Antimicrobial Agents and Chemotherapy* (2025). URL: https://inria.hal.science/hal-05444001. In press (cit. on p. 30).

[11] S. Dubois, M. Zytnicki, C. Lemaitre and T. Faraut. 'Pairwise graph edit distance characterizes the impact of the construction method on pangenome graphs'. In: *Bioinformatics* 41.6 (2nd June 2025), btaf291. DOI: 10.1101/2024.12.06.627166. URL: https://inria.hal.science/hal-04871087 (cit. on p. 22).

[12] C. Falentin, W. J.W. Thomas, M. Boudet, P. Le Boulch, A. Bourdais, L. Maillet, F. Legeai, K. Labadie, C. Cruaud, G. Deniot, J. Batley, A. Gravot, J.-M. Aury and M. Rousseau-Gueutin. 'Chromosome level assembly of five Brassica rapa and oleracea accessions expand the resistance genes reservoir'. In: *Scientific Data* 12.1 (2025), p. 2016. DOI: 10.1038/s41597-025-06261-5. URL: https://hal.science/hal-05421264 (cit. on p. 30).

[13] S. Gimenez, M. Eychenne, F. Legeai, S. Gamble and E. d'Alençon. 'Towards identification of a holocentromere marker in the lepidopteran model Spodoptera frugiperda'. In: *Chromosoma Biology of the Nucleus* 134.1 (11th Mar. 2025), p. 2. DOI: 10.1007/s00412-025-00828-2. URL: https://hal.inrae.fr/hal-05039973 (cit. on p. 30).

[14] N. Maurice, C. Lemaitre, R. Vicedomini and C. Frioux. 'Mapler: a pipeline for assessing assembly quality in taxonomically rich metagenomes sequenced with HiFi reads'. In: *Bioinformatics* 41.6 (6th June 2025), btaf334. DOI: 10.1093/bioinformatics/btaf334. URL: https://hal.science/hal-05288241 (cit. on p. 23).

[15] F. de Moor, O. Boullé and D. Lavenier. 'De-Bruijn graph partitioning for scalable and accurate DNA storage processing'. In: *Bioinformatics* 41.11 (1st Nov. 2025), pp. 1–9. DOI: 10.1093/bioinformatics/btaf618. URL: https://hal.science/hal-05375444 (cit. on p. 25).

[16] C. Moraga, C. Branco, Q. Rougemont, P. Veltsos, P. Jedlička, A. Muyle, M. Hanique, E. Tannier, X. Liu, E. Mendoza-Galindo et al. 'The Silene latifolia genome and its giant Y chromosome'. In: *Science* 387.6734 (6th Feb. 2025), pp. 630–636. DOI: 10.1126/science.adj7430. URL: https://hal.science/hal-04293712 (cit. on p. 29).

[17] S. G. Olvera-Vazquez, X. Chen, A. Mesnil, C. Meslin, F. Almeida-Silva, J. Confais, Y. Bourgeois, G. Lombardi, C. Lougmani, K. Alix, N. Francillonne, N. Choisne, S. Cauet, J.-C. Simon, C. Buchard, N. Rodde, D. Ogereau, C. Mottet, A. Degrave, E. Segura, A. Carbone, B. Benoît, E. Jacquin-Joly, W. Marande, D. Lavenier, F. Legeai and A. Cornille. 'Comprehensive annotation of olfactory and gustatory receptor genes and transposable elements revealed their evolutionary dynamics in aphids'. In: *Molecular Biology and Evolution* 42.12 (23rd Sept. 2025), msaf238. DOI: 10.1093/molbev/msaf238. URL: https://anses.hal.science/anses-05299488 (cit. on p. 29).

[18] C. Perrier, R. Allio, F. Legeai, M. Gautier, F. Bénéluz, W. Marande, A. Théron, N. Rodde, M. Herrera, L. Sauné, H. Parrinello, M. Mcclure and M. Arias. 'Transposable element accumulation drives genome size increase in Hylesia metabus (Lepidoptera: Saturniidae), an urticating moth species from South America'. In: *Journal of Heredity* 116.3 (May 2025), pp. 344–353. DOI: 10.1093/jhered/esae069. URL: https://cnrs.hal.science/hal-04792049 (cit. on p. 30).

[19] O. Sladký, P. Veselý and K. Břinda. 'FroM Superstring to Indexing: a space-efficient index for unconstrained k-mer sets using the Masked Burrows-Wheeler Transform (MBWT)'. In: *Bioinformatics Advances* (Nov. 2025), pp. 1–10. DOI: 10.1093/bioadv/vbaf290. URL: https://inria.hal.science/hal-04764171 (cit. on p. 20).

[20] T. Vericel, G. Gong, F. Legeai, A. Etier, J. Jaquiéry and J.-C. Simon. 'Life Cycle Simplifications in Aphids Drive Changes in Evolutionary Rates and Selection Regimes'. In: *Molecular Biology and Evolution* 42.12 (1st Dec. 2025), msaf307. DOI: 10.1093/molbev/msaf307. URL: https://hal.science/hal-05421198 (cit. on p. 30).

[21] R. Vicedomini, F. Andreace, Y. Dufresne, R. Chikhi and C. Duitama González. 'MUSET: set of utilities for constructing abundance unitig matrices from sequencing data [sequence analysis]'. In: *Bioinformatics* 41.3 (4th Mar. 2025), pp. 1–5. DOI: 10.1093/bioinformatics/btaf054. URL: https://hal.science/hal-04831168 (cit. on p. 24).

International peer-reviewed conferences

[22] A. Carroll, L. Mortimer, H. Ghosh, S. Reuter, H. Grundmann, K. Brinda, W. Hanage, A. Li, A. Purssell, B. Coburn, A. Rooney, S. Able-Thomas, M. Antonio, D. Macfadden and A. Mcgeer. 'P-397. Rapid Prediction of Genetic Relatedness of Escherichia coli Direct from Critical Care Samples Using Metagenomic Sequencing Paired with Neighbour Typing'. In: *Open Forum Infectious Diseases*. IDWeek 2024. Vol. 12. IDWeek 2024 Abstracts Supplement_1. Los Angeles, CA, United States: Oxford University Press, 2025, S357. DOI: 10.1093/ofid/ofae631.598. URL: https://hal.science/hal-05045814 (cit. on p. 30).

[23] M. Pegard, A. Poublan-Couzardot, P. Barre, S. de Givry, C. Gaspin, F. Legeai, F. Choulet, C. Klopp and B. Julier. 'Enhanced Genome Assemblies of French-Bred Dactylis glomerata and Medicago sativa: Achieving High-Quality Tetraploid Genomes'. In: *Breeding and Genetic Improvement for a Net-Zero Future*. FCAG 2025 - 36th Meeting of the EUCARPIA Fodder Crops and Amenity Grasses Section. Aberystwyth, United Kingdom, Sept. 2025, pp. 1–2. URL: https://hal.inrae.fr/hal-05265025 (cit. on p. 30).

[24] O. Sladký, P. Veselý and K. Břinda. 'Towards Efficient k-Mer Set Operations via Function-Assigned Masked Superstrings'. In: *Proceedings of the Prague Stringology Conference 2025*. PSC 2025 - Prague Stringology Conference. Prague, Czech Republic, Aug. 2025, pp. 26–40. DOI: 10.1101/2024.03.06.583483. URL: https://hal.science/hal-04573444 (cit. on p. 20).

Conferences without proceedings

[25] L. Ackermann, P. Peterlongo and K. Břinda. 'Towards space-efficient data structures for large genome-distance matrices with quick retrieval'. In: DSB 2025 - Workshop Data Structures in Bioinformatics. Pisa, Italy, 2025. URL: https://hal.science/hal-05418574 (cit. on pp. 13, 22).

[26] C. Airault, C. Deltel, F. de Moor, E. Drezen, M. Mognol and D. Lavenier. 'Protein database search using Processing-in-Memory architecture'. In: IPDPS 2025 - IEEE International Parallel and Distributed Processing Symposium. Milano, Italy: IEEE, 2025, pp. 939–948. DOI: 10.1109/IPDPSW66978.2025.00146. URL: https://hal.science/hal-05375748 (cit. on p. 26).

[27] L. Angevin, J. Lipovac, R. Vicedomini, P. Peterlongo and E. Roux. 'Identification of bacterial strains in gut microbiota: tools comparison and application'. In: Journées du PEPI IBIS 2025. Rennes, France, 2025, pp. 1–18. URL: https://hal.science/hal-05384947 (cit. on p. 28).

[28] L. Angevin, P. Peterlongo, R. Vicedomini, A. Siegel, J. Got and E. Roux. 'Metagenomic taxonomic assignment using Nanopore reads, reconstruction of metabolic networks and prediction of metabolite production'. In: 2025 - Journées Nutrition et Ecosystèmes Microbiens. Rennes, France, 2025. URL: https://hal.science/hal-05384891 (cit. on p. 27).

[29] F. Benoit, M. Temperville, M. Le Goff, C. Lemaitre, S. Glémin and C. Mérot. 'Detecting chromosomal inversions for population genomics: what could be the optimal approach?' In: GRC 2025 - Gordon Research Conference on Ecological and Evolutionary Genomics. Lucques (Barga), Italy, 2025. URL: https://hal.science/hal-05415935 (cit. on p. 27).

[30] F. Benoit, M. Temperville, M. Le Goff, C. Lemaitre, S. Glémin and C. Mérot. 'Detecting chromosomal inversions for population genomics: what could be the optimal approach?' In: Alphy & AIEM 2025 - ALignement et PHYlogénie - Approche Interdisciplinaire de l'Evolution Moléculaire. Lyon, France, 2025. URL: https://hal.science/hal-05411358 (cit. on p. 27).

[31] F. Brunetti and K. Brinda. 'Optimized K-mer Matching For Million-Genome Collections On Laptops'. In: DSB 2025 - Workshop Data Structures in Bioinformatics. Pisa, Italy, 5th Mar. 2025. URL: https://inria.hal.science/hal-05448274 (cit. on pp. 13, 21).

[32] F. Brunetti and K. Brinda. 'Rapid searches across million-genome bacterial collections on laptops: a practical methodology for point-of-care applications using k-mers'. In: SIM 2025 - 53rd National Congress of the Italian Society of Microbiology. Catania, Italy, 2025. URL: https://inria.hal.science/hal-05446703 (cit. on pp. 13, 21).

[33] N. Maurice, C. Lemaitre, C. Frioux and R. Vicedomini. 'Investigating pre-assembly clustering of HiFi reads for de novo assembly of complex metagenomes'. In: SeqBIM. Nantes, France, 24th Nov. 2025. URL: https://hal.science/hal-05444605 (cit. on p. 24).

[34] M. Mognol, F. de Moor, E. Drezen, Y. Falevoz and D. Lavenier. 'Evaluating Energy Efficiency of Genomics Algorithms on Processing-in-Memory Architectures'. In: PECS 2025 - International Workshop on Performance and Energy Efficiency in Concurrent and Distributed Systems. Dresden, Germany, 26th Aug. 2025, pp. 1–12. URL: https://hal.science/hal-05375465 (cit. on p. 26).

[35] A. Regnier and P. Peterlongo. 'k-mer matrix compression'. In: DSB 2025 Pisa - Workshop Data Structures in Bioinformatics. Pise, Italy, 2025. URL: https://hal.science/hal-05448857 (cit. on p. 21).

[36] O. Sladký, P. Veselý and K. Brinda. 'Masked superstrings as a compact, indexable, and dynamic representation of unconstrained k-mer sets'. In: DSB 2025 - Workshop Data Structures in Bioinformatics. Pisa, Italy, 2025. URL: https://inria.hal.science/hal-05448264 (cit. on p. 20).

[37] M. Temperville, F. Benoit, C. Mérot, F. Legeai and C. Lemaitre. 'SVJedi-Tag : a novel method for genotyping large inversions with linked-read data'. In: JOBIM 2025 - Journées Ouvertes Biologie, Informatique et Mathématiques. Bordeaux, France, 2025, pp. 1–9. URL: https://hal.science/hal-05393269 (cit. on p. 23).

[38] T. K. M. Truong, D. Lavenier, P. Peterlongo and K. Břinda. 'Optimized phylogenetic batching of million-genome collections for reduced storage requirements and faster data retrieval'. In: SeqBIM 2025. Nantes, France, 2025. URL: https://inria.hal.science/hal-05448733 (cit. on pp. 13, 22).

[39] R. Vicedomini and R. Chikhi. 'Improved strain-level metagenome assembly for modern long reads'. In: SeqBIM 2025 - Journées sur les Séquences en Bioinformatique, Informatique et Mathématiques. Nantes, France, 2025. URL: https://hal.science/hal-05424057 (cit. on p. 24).

Edition (books, proceedings, special issue of a journal)

[40] *SeqBIM 2025 Abstracts and Proceedings*. SeqBIM 2025. Nantes, France, 2025, pp. 1–30. URL: https://hal.science/hal-05412009 (cit. on p. 38).

Doctoral dissertations and habilitation theses

[41] M. Mognol. 'Acceleration of bioinformatics algorithms on a Processing-in-Memory architecture'. Université de Rennes, 2nd July 2025. URL: https://theses.hal.science/tel-05398591 (cit. on p. 41).

Reports & preprints

[42] C. Belliardo, N. Maurice, A. Pere, S. Mondy, A. Franc, M. Bailly-Bechet, C. Lemaitre, R. Vicedomini, J.-M. Frigerio, F. Salin, É. Belmonte, D. J. Sherman, P. Abad, C. Frioux and É. G. Danchin. *Accurate MAG reconstruction from complex soil microbiome through combined short- and HiFi long-reads metagenomics*. 12th Sept. 2025. DOI: 10.1101/2025.09.12.675765. URL: https://inria.hal.science/hal-05340126 (cit. on pp. 24, 28).

[43] S. Bocs, C. Carrette, J. Confais, S. Dubois, L. Duvaux, C. Klopp, N. Lapalu, P. Lasserre-Zuber, F. Legeai, C. Lemaitre, B. Linard, N. Marthe, B. Pierre, G. Sarah, F. Sabot, C. Tranchant-Dubreuil and M. Zytnicki. *A roadmap for the adoption of pangenomics in agronomy*. 2025. DOI: 10.5802/fake.doi. URL: https://hal.inrae.fr/hal-05357866 (cit. on p. 27).

[44] F. Brunetti and K. Břinda. *Optimized k-mer search across millions of bacterial genomes on laptops*. 26th Nov. 2025. DOI: 10.1101/2025.11.23.690050. URL: https://inria.hal.science/hal-05387449 (cit. on pp. 13, 21).

[45] M. Budia-Silva, A. Carroll, H. Ghosh, A. Mcgeer, T. Giani, G. M. Rossolini, K. Brinda, W. Hanage, H. Grundmann, D. Macfadden and S. Reuter. *Neighbour Typing Using Long Read Sequencing Provides Rapid Prediction of Sequence Type and Antimicrobial Susceptibility of Klebsiella pneumoniae*. 9th Sept. 2025. DOI: 10.1101/2025.09.03.673989. URL: https://inria.hal.science/hal-05263562 (cit. on p. 31).

[46] R. Chikhi, T. Lemane, R. Loll-Krippleber, M. Montoliu-Nerin, B. Raffestin, A. P. Camargo, C. J. Miller, M. B. Fiamenghi, D. P. Agustinho, S. Majidian, G. Autric, M. Hugues, J. Lee, R. Faure, K. D. Curry, J. A. Moura de Sousa, E. P. C. Rocha, D. Koslicki, P. Medvedev, P. Gupta, J. Shen, A. Morales-Tapia, K. Sihuta, P. J. Roy, G. W. Brown, R. C. Edgar, A. Korobeynikov, M. Steinegger, C. A. Lareau, P. Peterlongo and A. Babaian. *Logan: Planetary-Scale Genome Assembly Surveys Life's Diversity*. 31st July 2025. DOI: 10.1101/2024.07.30.605881. URL: https://inria.hal.science/hal-05446815 (cit. on p. 19).

[47] A. Kaul, F. W. Rossine, K. Brinda and M. Baym. *Novel genes arise from genomic deletions across the bacterial tree of life*. 5th Jan. 2026. DOI: 10.64898/2026.01.05.697752. URL: https://inria.hal.science/hal-05443487 (cit. on p. 28).

[48] V. Levallois, Y. Shibuya, B. Le Gal, R. Patro, P. Peterlongo and G. E. Pibiri. *Kaminari: a resource-frugal index for approximate colored k -mer queries*. 21st May 2025. DOI: 10.1101/2025.05.16.654317. URL: https://inria.hal.science/hal-05395000 (cit. on p. 20).

[49] J. Lipovac, M. Šikić, R. Vicedomini and K. Križanović. *MADRe: Strain-Level Metagenomic Classification Through Assembly-Driven Database Reduction*. 15th May 2025. DOI: 10.1101/2025.05.12.653324. URL: https://hal.science/hal-05343265 (cit. on pp. 25, 28).

[50] S. Romain, S. Dubois, F. Legeai and C. Lemaitre. *Investigating the topological motifs of inversions in pangenome graphs*. 17th Mar. 2025. DOI: 10.1101/2025.03.14.643331. URL: https://inria.hal.science/hal-05204329 (cit. on p. 27).

Other scientific publications

[51] L. Angevin, J. Lipovac, R. Vicedomini, P. Peterlongo and E. Roux. 'From bacterial taxonomic classification at strain level to metabolic networks producibility in gut microbiota'. In: 2025 - 14th International Gut Microbiology Symposium. Clermont-Ferrand, France, 2025. URL: https://hal.science/hal-05381840 (cit. on p. 27).

[52] C. Belliardo, A. Péré, A. Franc, J.-M. Frigerio, N. Maurice, C. Frioux, C. Lemaitre, S. Mondy, M. Bailly-Bechet, P. Abad, D. J. Sherman and E. G. Danchin. 'Enhancing Microbial Genome Reconstruction in Complex Environments by combining Short-and Long-read Sequencing'. In: 2025 - Journées scientifiques Agroécologie et Numérique. Dijon, France, 29th Jan. 2025, pp. 1–1. URL: https://hal.science/hal-04920790 (cit. on p. 24).

[53] F. Benoit, M. Temperville, C. Lemaitre and C. Mérot. 'Evaluating sequencing techniques for chromosomal inversion detection in non-model species: insights from Coelopa frigida genomes.' In: GRS 2025 - Ecological and Evolutionary Genomics): Elucidating the Evolutionary Dynamics of Adaptation in Fluctuating Environments. Lucques, Italy, 2025. URL: https://hal.science/hal-05411334 (cit. on p. 27).

[54] T. Brazier, C. Lemaitre and C. Mérot. 'Diversity of genomic structural variation across the Tree of Life'. In: Jacques Monod Speciation: "A multidimensional view of speciation: bridging micro and macro-evolution". Roscoff, France, 20th Oct. 2025. URL: https://hal.science/hal-05411846 (cit. on p. 29).

[55] T. Brazier, C. Lemaitre and C. Mérot. 'Structural genetic diversity across the Tree of Life: development of a long-read based pipeline for robust detection of SVs'. In: PopGroup58. Sheffield, United Kingdom, 6th Jan. 2025. URL: https://hal.science/hal-05411814 (cit. on p. 29).

[56] T. Brazier, C. Lemaitre and C. Mérot. 'Structural genetic diversity across the Tree of Life: development of a long-read based pipeline for robust detection of SVs'. In: ALPHY. Lyon, France, 3rd Feb. 2025. URL: https://hal.science/hal-05411819 (cit. on p. 29).

[57] M. Budia-Silva, A. C. Carroll, H. Ghosh, A. Mcgeer, T. Giani, G. Maria Rossolini, K. Brinda, W. P. Hanage, H. Grundmann, D. R. Macfadden and S. Reuter. 'Neighbour Typing Using Long Read Sequencing Provides Rapid Prediction of Sequence Type and Phenotype in Klebsiella pneumoniae sensu lato from Europe'. In: Applied Bioinformatics & Public Health Microbiology. Hinxton, Cambridge, United Kingdom, May 2025. URL: https://inria.hal.science/hal-05447485 (cit. on p. 31).

[58] N. Maurice, C. Frioux, C. Lemaitre and R. Vicedomini. 'Mapler: Assessing assembly quality in taxonomically-rich metagenomes sequenced with HiFi reads'. In: 2025 - Journées scientifiques Agroécologie et Numérique. Dijon, France, 2025, pp. 1–1. URL: https://hal.science/hal-04941137 (cit. on p. 23).

[59] O. Sladký, P. Veselý and K. Břinda. 'Masked Superstrings as a compact, indexable, and dynamic representation of unconstrained k-mer sets'. In: RECOMB 2025 - 29th International Conference of Research in Computational Molecular Biology. Seoul, Korea, South Korea, 2025, pp. 1–1. URL: https://hal.science/hal-05044964 (cit. on p. 20).

Scientific popularization

[60] D. Lavenier. *Stocker nos données sur ADN*. Espace de sciences, Rennes. 5th Mar. 2025, https://www.editions-apogee.com/espace-des-sciences/719-stocker-nos-dones-sur\bibrangedashadn.html. URL: https://hal.science/hal-05375812 (cit. on p. 42).

[61] D. Lavenier and M. Antonini. 'Stockage d'information sur ADN'. In: *Le calcul à découvert*. CNRS Editions, 23rd Jan. 2025. URL: https://hal.science/hal-05376118 (cit. on p. 42).

[62] E. Roux and G. Boudry. 'Biodiversité alimentaire, microbiote et bien-être : la recherche explore les liens potentiels'. In: *The Conversation France* (Nov. 2025). DOI: 10.64628/AAK.mux7urftw. URL: https://hal.science/hal-05369369 (cit. on p. 42).

12.3 Cited publications

[63] M. Alonge, X. Wang, M. Benoit, S. Soyk, L. Pereira, L. Zhang et al. 'Major Impacts of Widespread Structural Variation on Gene Expression and Crop Improvement in Tomato'. In: *Cell* (June 2020). DOI: 10.1016/j.cell.2020.05.021 (cit. on p. 10).

[64] S. F. Altschul, W. Gish, W. Miller, E. W. Myers and D. J. Lipman. 'Basic local alignment search tool'. In: *Journal of molecular biology* 215.3 (1990), pp. 403–410 (cit. on p. 8).

[65] G. Benoit, M. Mariadassou, S. Robin, S. Schbath, P. Peterlongo and C. Lemaitre. 'SimkaMin: fast and resource frugal de novo comparative metagenomics'. In: *Bioinformatics* 36.4 (2020), pp. 1275–1276 (cit. on p. 12).

[66] G. Benoit, P. Peterlongo, M. Mariadassou, E. Drezen, S. Schbath, D. Lavenier and C. Lemaitre. 'Multiple comparative metagenomics using multiset k-mer counting'. In: *PeerJ Computer Science* 2 (2016), e94 (cit. on p. 12).

[67] Y. Dong, F. Sun, Z. Ping, Q. Ouyang and L. Qian. 'DNA storage: research landscape and future prospects'. In: *National Science Review* 7.6 (Jan. 2020), pp. 1092–1107 (cit. on p. 11).

[68] A. Doricchi, C. M. Platnich, A. Gimpel, F. Horn, M. Earle, G. Lanzavecchia, A. L. Cortajarena, L. M. Liz-Marzán, N. Liu, R. Heckel, R. N. Grass, R. Krahne, U. F. Keyser and D. Garoli. 'Emerging Approaches to DNA Data Storage: Challenges and Prospects'. In: *ACS Nano* 16.11 (2022), pp. 17552–17571 (cit. on p. 11).

[69] E. Drezen, G. Rizk, R. Chikhi, C. Deltel, C. Lemaitre, P. Peterlongo and D. Lavenier. 'GATB: genome assembly & analysis tool box'. In: *Bioinformatics* 30.20 (2014), pp. 2959–2961 (cit. on p. 12).

[70] Y. Dufresne, T. Lemane, P. Marijon, P. Peterlongo, A. Rahman, M. Kokot, P. Medvedev, S. Deorowicz and R. Chikhi. 'The K-mer File Format: a standardized and compact disk representation of sets of k-mers'. In: *Bioinformatics* 38.18 (Sept. 2022), pp. 4423–4425. DOI: 10.1093/bioinformatics/btac528. URL: https://inria.hal.science/hal-03885245 (cit. on p. 12).

[71] S. Fatumo, A. Yakubu, O. Oyedele, J. Popoola, D. A. Attipoe, G. Eze-Echesi, F. Z. Modibbo, N. Ado-Wanka et al. 'Promoting the genomic revolution in Africa through the Nigerian 100K Genome Project'. In: *Nature Genetics* 54.5 (2022), pp. 531–536 (cit. on p. 6).

[72] D. Fujiki, X. Wang, A. Subramaniyan and R. Das. *In-/near-memory Computing*. Springer, 2021 (cit. on p. 11).

[73] E. Garrison, A. Guarracino, S. Heumos, F. Villani, Z. Bao, L. Tattini, J. Hagmann, S. Vorbrugg, S. Marco-Sola, C. Kubica, D. G. Ashbrook, K. Thorell, R. L. Rusholme-Pilcher, G. Liti, E. Rudbeck, S. Nahnsen, Z. Yang, M. N. Moses, F. L. Nobrega, Y. Wu, H. Chen, J. de Ligt, P. H. Sudmant, N. Soranzo, V. Colonna, R. W. Williams and P. Prins. 'Building pangenome graphs'. In: *bioRxiv* (Apr. 2023). DOI: 10.1101/2023.04.05.535718 (cit. on p. 10).

[74] J. Gauthier, D. L. de Silva, Z. Gompert, A. Whibley, C. Houssin, Y. Le Poul, M. McClure, C. Lemaitre, F. Legeai, J. Mallet and M. Elias. 'Contrasting genomic and phenotypic outcomes of hybridization between pairs of mimetic butterfly taxa across a suture zone'. In: *Molecular Ecology* 29.7 (Apr. 2020), pp. 1328–1343. DOI: 10.1111/mec.15403 (cit. on p. 12).

[75] J. S. Ghurye, V. Cepeda-Espinoza and M. Pop. 'Metagenomic Assembly: Overview, Challenges and Applications'. In: *The Yale Journal of Biology and Medicine* 89.3 (2016), p. 353 (cit. on p. 10).

[76] C. Guyomar, F. Legeai, E. Jousselin, C. Mougel, C. Lemaitre and J.-C. Simon. 'Multi-scale characterization of symbiont diversity in the pea aphid complex through metagenomic approaches'. In: *Microbiome* 6.1 (Oct. 2018), p. 181. DOI: 10.1186/s40168-018-0562-9 (cit. on p. 12).

[77] G. Hickey, J. Monlong, J. Ebler, A. M. Novak, J. M. Eizenga, Y. Gao, H. J. Abel, L. L. Antonacci-Fulton, M. Asri et al. 'Pangenome graph construction from genome alignments with Minigraph-Cactus'. In: *Nature Biotechnology* (May 2023). DOI: 10.1038/s41587-023-01793-w (cit. on p. 10).

[78] M. Hunt, L. Lima, D. Anderson, G. Bouras, M. Hall, J. Hawkey, O. Schwengers, W. Shen, J. A. Lees and Z. Iqbal. 'AllTheBacteria–all bacterial genomes assembled, available, and searchable'. In: *BioRxiv* (2024), pp. 2024–03 (cit. on pp. 6, 13).

[79] M. Karasikov, H. Mustafa, D. Danciu, O. Kulkov, M. Zimmermann, C. Barber, G. Rätsch and A. Kahles. 'Efficient and accurate search in petabase-scale sequence repositories'. In: *Nature* (2025), pp. 1–9 (cit. on p. 9).

[80] K. Katz, O. Shutov, R. Lapoint, M. Kimelman, J. R. Brister and C. O'Sullivan. 'The sequence read archive: a decade more of explosive growth'. In: *Nucleic acids research* 50.D1 (2022), pp. D387–D390 (cit. on p. 9).

[81] D. Lavenier, J.-F. Roy and D. Furodet. 'DNA Mapping using Processor-in-Memory Architecture'. In: *Workshop on Accelerator-Enabled Algorithms and Applications in Bioinformatics*. Shenzhen, China, Dec. 2016. URL: https://hal.science/hal-01399997 (cit. on p. 11).

[82] F. Legeai, B. F. Santos, S. Robin, A. Bretaudeau, R. B. Dikow, C. Lemaitre, V. Jouan, M. Ravallec, J.-M. Drezen, D. Tagu, F. Baudat, G. Gyapay, X. Zhou, S. Liu, B. A. Webb, S. G. Brady and A.-N. Volkoff. 'Genomic architecture of endogenous ichnoviruses reveals distinct evolutionary pathways leading to virus domestication in parasitic wasps'. In: *BMC Biology* 18.1 (July 2020). DOI: 10.1186/s12915-020-00822-3 (cit. on p. 12).

[83] T. Lemane, N. Lezzoche, J. Lecubin, E. Pelletier, M. Lescot, R. Chikhi and P. Peterlongo. 'Indexing and real-time user-friendly queries in terabyte-sized complex genomic datasets with kmindex and ORA'. In: *Nature Computational Science* 4.2 (2024), pp. 104–109 (cit. on pp. 12, 19).

[84] T. Lemane, N. Lezzoche, J. Lecubin, E. Pelletier, M. Lescot, R. Chikhi and P. Peterlongo. 'Indexing and real-time user-friendly queries in terabyte-sized complex genomic datasets with kmindex and ORA'. In: *Nature Computational Science* 4.2 (2024), pp. 104–109 (cit. on p. 21).

[85] T. Lemane, P. Medvedev, R. Chikhi and P. Peterlongo. 'kmtricks: efficient and flexible construction of Bloom filters for large sequencing data collections'. In: *Bioinformatics Advances* 2.1 (Jan. 2022). Ed. by T. Lengauer. DOI: 10.1093/bioadv/vbac029. URL: http://dx.doi.org/10.1093/bioadv/vbac029 (cit. on p. 19).

[86] W.-W. Liao, M. Asri, J. Ebler, D. Doerr, M. Haukness, G. Hickey, S. Lu, J. K. Lucas, J. Monlong, H. J. Abel et al. 'A draft human pangenome reference'. In: *Nature* 617.7960 (2023), pp. 312–324 (cit. on p. 6).

[87] W.-W. Liao, M. Asri, J. Ebler, D. Doerr, M. Haukness, G. Hickey, S. Lu, J. K. Lucas, J. Monlong, H. J. Abel et al. 'A draft human pangenome reference'. In: *Nature* 617.7960 (May 2023), pp. 312–324. DOI: 10.1038/s41586-023-05896-x (cit. on p. 10).

[88] A. Limasset, G. Rizk, R. Chikhi and P. Peterlongo. 'Fast and scalable minimal perfect hashing for massive key sets'. In: *16th International Symposium on Experimental Algorithms*. Vol. 11. 2017, pp. 1–11 (cit. on p. 12).

[89] L. Luo, D. Guo, R. Ma, O. Rottenstreich and X. Luo. 'Optimizing Bloom Filter: Challenges, Solutions, and Comparisons'. In: *IEEE Communications Surveys & Tutorials* PP (Apr. 2018). DOI: 10.1109/COMST.2018.2889329 (cit. on p. 11).

[90] M. Mahmoud, N. Gobet, D. I. Cruz-Dávalos, N. Mounier, C. Dessimoz and F. J. Sedlazeck. 'Structural variant calling: the long and the short of it'. In: *Genome Biology* 20.1 (Nov. 2019) (cit. on p. 10).

[91] D. Molka, D. Hackenberg, R. Schone and M. S. Muller. 'Characterizing the energy consumption of data transfers and arithmetic operations on x86-64 processors'. In: *International conference on green computing*. USA: IEEE Computer Society, 2010. DOI: 10.1109/GREENCOMP.2010.5598316. URL: https://doi.org/10.1109/GREENCOMP.2010.5598316 (cit. on p. 11).

[92] P. Morisse, C. Lemaitre and F. Legeai. 'LRez: a C++ API and toolkit for analyzing and managing Linked-Reads data'. In: *Bioinformatics Advances* 1.1 (Jan. 2021). DOI: 10.1093/bioadv/vbab022 (cit. on p. 12).

[93] B. D. Ondov, T. J. Treangen, P. Melsted, A. B. Mallonee, N. H. Bergman, S. Koren and A. M. Phillippy. 'Mash: fast genome and metagenome distance estimation using MinHash'. In: *Genome biology* 17.1 (2016), pp. 1–14 (cit. on pp. 12, 22).

[94] M. H. Raza, S. Desai, S. Aravamudhan and R. Zadegan. 'An outlook on the current challenges and opportunities in DNA data storage'. In: *Biotechnology Advances* 66 (2023), p. 108155. DOI: https://doi.org/10.1016/j.biotechadv.2023.108155. URL: https://www.sciencedirect.com/science/article/pii/S0734975023000629 (cit. on p. 11).

[95] G. Rizk, A. Gouin, R. Chikhi and C. Lemaitre. 'MindTheGap : integrated detection and assembly of short and long insertions.' In: *Bioinformatics* 30.24 (Dec. 2014), pp. 3451–3457. DOI: 10.1093/bioinformatics/btu545 (cit. on p. 12).

[96] S. Romain and C. Lemaitre. 'SVJedi-graph: improving the genotyping of close and overlapping structural variants with long reads using a variation graph'. In: *Bioinformatics* 39.Supplement 1 (June 2023), pp. i270–i278. DOI: 10.1093/bioinformatics/btad237 (cit. on p. 12).

[97] S. A. Shiryev and R. Agarwala. 'Indexing and searching petabyte-scale nucleotide resources'. In: *bioRxiv* (2023), pp. 2023–07 (cit. on p. 12).

[98] G. Siekaniec, E. Roux, T. Lemane, E. Guédon and J. Nicolas. 'Identification of isolated or mixed strains from long reads: a challenge met on Streptococcus thermophilus using a MinION sequencer'. In: *Microbial genomics* 7.11 (2021), p. 000654 (cit. on p. 28).

[99] O. Sladký, P. Veselý and K. Břinda. 'Masked superstrings as a unified framework for textual k-mer set representations'. In: *bioRxiv* 2023.02.01.526717 (2023). DOI: 10.1101/2023.02.01.526717 (cit. on p. 20).

[100] S. Sunagawa, S. G. Acinas, P. Bork, C. Bowler, D. Eveillard, G. Gorsky, L. Guidi, D. Iudicone, E. Karsenti et al. 'Tara Oceans: towards global ocean ecosystems biology'. In: *Nature Reviews Microbiology* 18.8 (2020), pp. 428–445 (cit. on p. 6).

[101] R. Uricaru, G. Rizk, V. Lacroix, E. Quillery, O. Plantard, R. Chikhi, C. Lemaitre and P. Peterlongo. 'Reference-free detection of isolated SNPs'. In: *Nucleic acids research* 43.2 (2015), e11–e11 (cit. on p. 12).

[102] R. Vicedomini, C. Quince, A. E. Darling and R. Chikhi. 'Strainberry: automated strain separation in low-complexity metagenomes using long reads'. In: *Nature Communications* 12.1 (2021), p. 4485 (cit. on p. 10).

[103] T. Wang, L. Antonacci-Fulton, K. Howe, H. A. Lawson, J. K. Lucas, A. M. Phillippy, A. B. Popejoy, M. Asri, C. Carson, M. J. Chaisson et al. 'The Human Pangenome Project: a global resource to map genomic diversity'. In: *Nature* 604.7906 (2022), pp. 437–446 (cit. on p. 6).

[104] J. Weischenfeldt, O. Symmons, F. Spitz and J. O. Korbel. 'Phenotypic impact of genomic structural variation: insights from and for human disease'. In: *Nature Reviews Genetics* 14.2 (Jan. 2013), pp. 125–138. DOI: 10.1038/nrg3373 (cit. on p. 10).

[105] M. Wellenreuther, C. Mérot, E. Berdan and L. Bernatchez. 'Going beyond SNPs: The role of structural genomic variants in adaptive evolution and species diversification'. In: *Molecular Ecology* 28.6 (Mar. 2019), pp. 1203–1209. DOI: 10.1111/mec.15066 (cit. on p. 10).