UMR IRISA

Activity Report 2024

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

Contents

Project-Team GENSCALE 1

1 Team members, visitors, external collaborators 2

2 Overall objectives 3

3 Research program 4
- 3.1 Axis 1: Data structures and indexing algorithms . . . 4
- 3.2 Axis 2: Sequence analysis algorithms . . . 5
- 3.3 Axis 3: Parallelism . . . 5
- 3.4 Axis 4: Applications . . . 6

4 Application domains 6
- 4.1 Health . . . 6
- 4.2 Agronomy . . . 6
- 4.3 Environment . . . 7

5 Social and environmental responsibility 7
- 5.1 Impact of research results . . . 7

6 Highlights of the year 7
- 6.1 Awards . . . 7
- 6.2 Results . . . 7

7 New software, platforms, open data 8
- 7.1 New software . . . 8
  - 7.1.1 kmtricks . . . 8
  - 7.1.2 kmindex . . . 8
  - 7.1.3 kmdiff . . . 8
  - 7.1.4 muset . . . 9
  - 7.1.5 cdbgtricks . . . 9
  - 7.1.6 back to sequences . . . 9
  - 7.1.7 KmerCamel . . . 9
  - 7.1.8 Phylign . . . 10
  - 7.1.9 MiniPhy . . . 10
  - 7.1.10 gfagraphs . . . 10
  - 7.1.11 pancat . . . 10
  - 7.1.12 rs-pancat-compare . . . 11
  - 7.1.13 Mapler . . . 11
  - 7.1.14 HairSplitter . . . 11
  - 7.1.15 Alice . . . 11
  - 7.1.16 DnarXiv . . . 12

8 New results 12
- 8.1 Data structures and indexing algorithms . . . 12
  - 8.1.1 FMSI k-mer indexing via masked superstrings . . . 12
  - 8.1.2 Dynamic k-mer set operations via masked superstrings . . . 12
  - 8.1.3 Optimized K-mer Matching For Million-Genome Collections On Laptops . . . 13
  - 8.1.4 Updating compacted de Bruijn graphs . . . 13
  - 8.1.5 The Backpack Quotient Filter: a dynamic and space-efficient data structure for querying k-mers with abundance . . . 13
  - 8.1.6 Improve the resizing of the Backpack Quotient Filter . . . 14
  - 8.1.7 Back to sequences: Find the origin of k-mers . . . 14
  - 8.1.8 Indexing and real-time user-friendly queries in terabyte-sized complex genomic datasets with kmindex and ORA. . . . 14

8.1.9 Logan Search, a k-mer search engine for all Sequence Read Archive public accessions 14
8.1.10 k-mer matrix compression . . . 15
8.1.11 Towards space-efficient data structures for large genome-distance matrices with quick retrieval . . . 15
8.2 Sequence analysis algorithms . . . 16
8.2.1 Haplotype assembly from long and noisy reads . . . 16
8.2.2 Scaffolding step in genome assembly . . . 16
8.2.3 Assessing assembly quality in metagenomes of increasing complexity sequenced with HiFi long reads . . . 16
8.2.4 Construction of unitig matrices with abundance information . . . 17
8.2.5 Pairwise graph edit distance characterizes the impact of the construction method on pangenome graphs . . . 17
8.2.6 Pangenome graph manipulation for local visualisation with pancat . . . 18
8.2.7 Understanding the limits of pangenome graphs for the analysis of large inversions . 18
8.3 Information storage on DNA molecules . . . 19
8.3.1 Design of long DNA molecules . . . 19
8.3.2 Automation of biotechnology protocols . . . 19
8.3.3 Random access with optimized primers . . . 19
8.3.4 DNA caching . . . 20
8.4 Processing-in-Memory . . . 20
8.4.1 Sorting . . . 20
8.4.2 Alignment of long reads . . . 20
8.4.3 Genome compression . . . 20
8.5 Applications and bioinformatics analyses . . . 21
8.5.1 Unlocking the Soil Microbiome: Unraveling Soil Microbial Complexity Using Long-Read Metagenomics. . . . 21
8.5.2 Chromosome-Level Assembly and Annotation of the Pearly Heath Coenonympha arcania Butterfly Genome . . . 21
8.5.3 Genomics and transcriptomics of Brassicaceae plants and agro-ecosystem insects . 21
8.5.4 Rapid diagnostics of antibiotic resistance . . . 22

9 Bilateral contracts and grants with industry 22

10 Partnerships and cooperations 23
10.1 International research visitors . . . 23
10.1.1 Visits of international scientists . . . 23
10.2 European initiatives . . . 24
10.2.1 H2020 projects . . . 24
10.3 National initiatives . . . 25
10.3.1 PEPR . . . 25
10.3.2 ANR . . . 27
10.3.3 Inria Exploratory Action . . . 29
10.4 Regional initiatives . . . 30

11 Dissemination 31
11.1 Promoting scientific activities . . . 31
11.1.1 Scientific events: organisation . . . 32
11.1.2 Scientific events: selection . . . 32
11.1.3 Journal . . . 32
11.1.4 Invited talks . . . 33
11.1.5 Leadership within the scientific community . . . 33
11.1.6 Scientific expertise . . . 33
11.1.7 Research administration . . . 33
11.2 Teaching - Supervision - Juries . . . 33
11.2.1 Teaching administration . . . 33

11.2.2 Teaching . . . 34
11.2.3 HDR defense . . . 34
11.2.4 Supervision . . . 34
11.2.5 Juries . . . 35
11.3 Popularization . . . 36
11.3.1 Specific official responsibilities in science outreach structures . . . 36
11.3.2 Productions (articles, videos, podcasts, serious games, ...) . . . 36
11.3.3 Participation in Live events . . . 36

12 Scientific production **36**
12.1 Major publications . . . 36
12.2 Publications of the year . . . 37
12.3 Cited publications . . . 40

Project-Team GENSCALE

Creation of the Project-Team: 2013 January 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

A3.1.1. – Modeling, representation

A3.1.2. – Data management, quering and storage

A3.1.8. – Big data (production, storage, transfer)

A3.3.3. – Big data analysis

A7.1. – Algorithms

A7.1.3. – Graph algorithms

A8.2. – Optimization

A9.6. – Decision support

Other research topics and application domains

B1.1.4. – Genetics and genomics

B1.1.7. – Bioinformatics

B2.2.6. – Neurodegenerative diseases

B3.5. – Agronomy

B3.6. – Ecology

B3.6.1. – Biodiversity

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, from Oct 2024, HDR]
- Claire Lemaitre [INRIA, Researcher, until Sep 2024, HDR]
- Jacques Nicolas [INRIA, Senior Researcher]
- Riccardo Vicedomini [CNRS, Researcher]

Faculty Members

- Rumen Andonov [UNIV RENNES I, Emeritus, from Sep 2024, HDR]
- Rumen Andonov [UNIV RENNES I, Professor, until Aug 2024, HDR]
- Khodor Hannoush [UNIV RENNES I, ATER, from Sep 2024]

Post-Doctoral Fellow

- Loren Dejoies [INRIA, Post-Doctoral Fellow]

PhD Students

- Léo Ackermann [INRIA, from Oct 2024]
- Lune Angevin [UNIV RENNES I, from Oct 2024]
- Siegfried Dubois [INRAE]
- Roland Faure [UNIV LIBRE BRUXELLES, from Oct 2024 until Nov 2024]
- Roland Faure [UNIV RENNES I, until Sep 2024]
- Khodor Hannoush [INRIA, until Aug 2024]
- Victor Levallois [INRIA]
- Nicolas Maurice [INRIA]
- Meven Mognol [UNIV RENNES, CIFRE]
- Alix Regnier [INRIA, from Sep 2024]
- Sandra Romain [INRIA, until Oct 2024]
- Melody Temperville [UNIV RENNES I, from Oct 2024]
- Khac Minh Tam Truong [UNIV RENNES I, from Nov 2024]

Technical Staff

- Charly Airault [CNRS, Engineer, from Mar 2024]
- Olivier Boulle [CNRS, Engineer, from Oct 2024]
- Olivier Boulle [INRIA, Engineer, until Sep 2024]
- Julien Leblanc [CNRS, Engineer]
- Florestan de Moor [CNRS, Engineer]

Interns and Apprentices

- Lune Angevin [INRAE, Intern, until Jul 2024]
- Nicolas Buchin [CNRS, Intern, from May 2024 until Jul 2024]
- Arya Kaul [INRIA, Intern, until Mar 2024]
- Alix Regnier [INRIA, Intern, from Feb 2024 until Jul 2024]
- Melody Temperville [INRIA, Intern, from Feb 2024 until Jun 2024]

Administrative Assistant

- Marie Le Roic [INRIA]

Visiting Scientists

- Francesca Brunetti [UNIV SAPIENZA , until Nov 2024]
- Veronika Hendrychová [Czech Technical University in Prague, until Jun 2024]
- Josipa Lipovac [UNIV ZAGREB, from Nov 2024]
- Ulysse Mc Connell [ETH Zurich, from Jun 2024 until Jul 2024]

External Collaborators

- Susete Alves Carvalho [INRAE, until Feb 2024]
- Erwan Drezen [INSTITUT PASTEUR]
- Fabrice Legeai [INRAE]
- Emeline Roux [UNIV RENNES I]

2 Overall objectives

The main goal of the GenScale project is to develop scalable methods and software programs for processing genomic data. Our research is motivated by the fast development of sequencing technologies, especially next-generation sequencing (NGS), and third-generation sequencing (TGS). NGS provides up to billions of very short (few hundreds of base pairs, bps) DNA fragments of high quality, called short reads, and TGS provides millions of long (thousands to millions of bps) DNA fragments of lower quality called long reads. Synthetic long reads or linked-reads is another technology type that combines the high quality and low cost of short-reads sequencing with long-range information by adding barcodes that tag reads originating from the same long DNA fragment. All these sequencing data bring very challenging problems both in terms of bioinformatics and computer science. As a matter of fact, the recent

sequencing machines generate terabytes of DNA sequences to which time-consuming processes must be applied to extract useful and relevant biological information.

A large panel of biological questions can be investigated using genomic data. A complete project includes DNA extraction from one or several living organisms, sequencing with high throughput machines, and finally the design of methods and development of bioinformatics pipelines to answer the initial question. Such pipelines are made of pre-processing steps (quality control and data cleaning), core functions transforming these data into genomic objects on which GenScale's expertise focuses (genome assembly, variant discovery -SNP, structural variations-, sequence annotation, sequence comparison, etc.) and sometimes further integration steps helping to interpret and gain some knowledge from data by incorporating other sources of semantic information.

The challenge for GenScale is to develop scaling algorithms able to devour the daily sequenced DNA flow that tends to congest the bioinformatics computing centers. To achieve this goal, our strategy is to work both on space and time scalability aspects. Space scalability is correlated to the design of optimized and low memory footprint data structures able to capture all useful information contained in sequencing datasets. The idea is to represent tera- or petabytes of raw data in a very concise way so that their analyses completely fit into a computer memory. Time scalability means that the execution of the algorithms must be linear with respect to the size of the problem or, at least, must last a reasonable amount of time. In this respect, parallelism is a complementary technique for increasing scalability.

A second important objective of GenScale is to create and maintain permanent partnerships with life science research groups. Collaboration with genomics research teams is of crucial importance for validating our tools, and for scientific watch in this extremely dynamic field. Our approach is to actively participate in solving biological problems (with our partners) and to get involved in a few challenging genomic projects.

GenScale research is organized along **four main axes**:

- Axis 1: Data structures & Indexing algorithms;
- Axis 2: Sequence analysis algorithms
- Axis 3: Parallelism
- Axis 4: Applications

3 Research program

3.1 Axis 1: Data structures and indexing algorithms

The aim of this axis is to create and diffuse efficient data structures for representing the mass of genomic data generated by the sequencing machines. This is necessary because the processing of large genomes, such as those of mammals or plants, or multiple genomes from a single sample in metagenomics, requires significant computing resources and a powerful memory configuration. The advances in TGS (Third Generation Sequencers) technologies bring also new challenges to represent or search information based on sequencing data with a high error rate.

Part of our research focuses on kmer representation (words of length k), and on the de-Bruijn graph structure. This well-known data structure, directly built from raw sequencing data, has many properties that match perfectly well with NGS processing requirements. Here, the question we are interested in is how to provide a low memory footprint implementation of the de-Bruijn graph to process very large NGS datasets, including metagenomic ones [4, 5].

A correlated research direction is the indexing of large sets of objects [8]. A typical, but non exclusive, need is to annotate nodes of the de-Bruijn graph, that is, potentially billions of items. Again, very low memory footprint indexing structures are mandatory to manage such a large quantity of objects [9].

3.2 Axis 2: Sequence analysis algorithms

The main goal of the GenScale team is to develop optimized tools dedicated to genomic data processing. Optimization can be seen both in terms of space (low memory footprint) and in terms of time (fast execution time). The first point is mainly related to advanced data structures as presented in the previous section (axis 1). The second point relies on new algorithms and, when possible, implementations on parallel structures (axis 3).

We do not have the ambition to cover the vast panel of software related to genomic data processing needs. We particularly focused on the following areas:

- **NGS data Compression** De-Bruijn graphs are *de facto* a compressed representation of the NGS information from which very efficient and specific compressors can be designed. Furthermore, compressing the data using smart structures may speed up some downstream graph-based analyses since a graph structure is already built [2].

- **Genome assembly** This task remains very complicated, especially for large and complex genomes, such as plant genomes with polyploid and highly repeated structures. We worked both on the generation of contigs [4] and on the scaffolding step [1]. Both NGS and TGS technologies are taken into consideration, either independently or using combined approaches.

- **Detection of variants** This is often the main information one wants to extract from the sequencing data. Variants range from SNPs or short indels to structural variants that are large insertions/deletions and long inversions over the chromosomes. We developed original methods to find variants without any reference genome [11], to detect structural variants using local NGS assembly approaches [10] or TGS processing.

- **Metagenomics** We focused our research on comparative metagenomics by providing methods able to compare hundreds of metagenomic samples together. This is achieved by combining very low memory data structures and efficient implementation and parallelization on large clusters [3].

- **Large scale indexation** We develop approaches, indexing terabyte sized datasets in a few days. As a result, those index make possible the query a sequence in a few minutes [8].

- **Storing information on DNA molecules** The DNA molecule can be seen as a promising support for information storage. This can be achieved by encoding information into DNA alphabet, including error correction codes and data security, before to synthetize the corresponding DNA molecules. Novel sequence algorithms need to be developed to take advantage of the specificities of these sequences [7].

3.3 Axis 3: Parallelism

This third axis investigates a supplementary way to increase performances and scalability of genomic treatments. There are many levels of parallelism that can be used and/or combined to reduce the execution time of very time-consuming bioinformatics processes. A first level is the parallel nature of today processors that now house several cores. A second level is the grid structure that is present in all bioinformatics centers or in the cloud. These two levels are generally combined: a node of a grid is often a multicore system. Another possibility is to work with processing in memory (PIM) boards or to add hardware accelerators to a processor. A GPU board is a good example.

GenScale does not do explicit research on parallelism. It exploits the capacity of computing resources to support parallelism. The problem is addressed in two different directions. The first is an engineering approach that uses existing parallel tools to implement algorithms such as multithreading or MapReduce techniques [5]. The second is a parallel algorithmic approach: during the development step, the algorithms are constrained by parallel criteria [3]. This is particularly true for parallel algorithms targeting hardware accelerators.

3.4 Axis 4: Applications

Sequencing data are intensively used in many life science projects. Thus, methodologies developed by the GenScale group are applied to a large panel of life sciences domains. Most of these applications face specific methodological issues that the team proposes to answer by developing new tools or by adapting existing ones. Such collaborations lead therefore to novel methodological developments that can be directly evaluated on real biological data and often lead to novel biological results. In most cases, we also participate in the data analyses and interpretations in terms of biological findings.

Furthermore, GenScale actively creates and maintains permanent partnerships with several local, national, or international groups, bearers of applications for the tools developed by the team and able to give valuable and relevant feedbacks.

4 Application domains

Today, sequencing data are intensively used in many life science projects. The methodologies developed by the GenScale group are generic approaches that can be applied to a large panel of domains such as health, agronomy or environment areas. The next sections briefly describe examples of our activity in these different domains.

4.1 Health

Genetic and cancer disease diagnostic: Genetic diseases are caused by some particular mutations in the genomes that alter important cell processes. Similarly, cancer comes from changes in the DNA molecules that alter cell behavior, causing uncontrollable growth and malignancy. Pointing out genes with mutations helps in identifying the disease and in prescribing the right drugs. Thus, DNA from individual patients is sequenced and the aim is to detect potential mutations that may be linked to the patient's disease. Bioinformatics analysis can be based on the detection of SNPs (Single Nucleotide Polymorphism) from a set of predefined target genes. One can also scan the complete genome and report all kinds of mutations, including complex mutations such as large insertions or deletions, that could be associated with genetic or cancer diseases.

Gut-brain axis and regulation of food behavior: Consumption of a Western-style diet has a major impact on the composition of the intestinal microbiota, and thus on the metabolites it produces. Diet-induced dysbiosis then triggers disturbances in specific gut functions, which in turn can influence the development of brain structures, including those involved in the regulation of food intake. Fine characterization of the intestinal microbiota, on a genome scale, gives access to the potential for metabolite production by the microbiota. This could lead to a better understanding of the impact of dysbiosis on the host, and to solutions to counteract this dysbiosis (probiotics, targeted fermented foods, metabolite supplementation, ...). The development of efficient methods, from metagenomic sequencing to genome production (de novo assembly and/or reference-based methods, ...), is a prerequisite for functional analysis.

4.2 Agronomy

Insect genomics: Insects represent major crop pests, justifying the need for control strategies to limit population outbreaks and the dissemination of plant viruses they frequently transmit. Several issues are investigated through the analysis and comparison of their genomes: understanding their phenotypic plasticity such as their reproduction mode changes, identifying the genomic sources of adaptation to their host plant and of ecological speciation, and understanding the relationships with their bacterial symbiotic communities [6].

Improving plant breeding: Such projects aim at identifying favorable alleles at loci contributing to phenotypic variation, characterizing polymorphism at the functional level and providing robust multi-locus SNP-based predictors of the breeding value of agronomical traits under polygenic control. Underlying bioinformatics processing is the detection of informative zones (QTL) on the plant genomes.

4.3 Environment

Food quality control: One way to check food contaminated with bacteria is to extract DNA from a product and identify the different strains it contains. This can now be done quickly with low-cost sequencing technologies such as the MinION sequencer from Oxford Nanopore Technologies.
Ocean biodiversity: The metagenomic analysis of seawater samples provides an original way to study the ecosystems of the oceans. Through the biodiversity analysis of different ocean spots, many biological questions can be addressed, such as the plankton biodiversity and its role, for example, in the CO2 sequestration.

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.

6 Highlights of the year

6.1 Awards

Roland Faure and Tam Khac Minh Truong were awarded the prize for the best student paper co-authored with Rumen Andonov at the BIOINFORMATICS 2024 conference in Rome [41].

6.2 Results

The indexing tool, kmindex, was published in Nature Computational Science [8]. This work enabled for the first time the indexing of hundreds of TB of raw metagenomics data for instant queries. Applied to Tara Oceans data, it enabled us to propose the first search engine able to perform queries on the full Tara Oceans Metagenomic dataset.

7 New software, platforms, open data

7.1 New software

7.1.1 kmtricks

Keywords: High throughput sequencing, Indexing, K-mer, Bloom filter, K-mers matrix

Functional Description: kmtricks is a tool suite built around the idea of k-mer matrices. It is designed for counting k-mers, and constructing bloom filters or counted k-mer matrices from large and numerous read sets. It takes as inputs sequencing data (fastq) and can output different kinds of matrices compatible with common k-mers indexing tools. The software is composed of several command line tools, a C++ library, and a C++ plugin system to extend its features.

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

Publication: hal-03166007

Contact: Pierre Peterlongo

Participants: Teo Lemane, Rayan Chikhi, Pierre Peterlongo

7.1.2 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.

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

Contact: Pierre Peterlongo

7.1.3 kmdiff

Keywords: K-mer, K-mers matrix, GWAS

Functional Description: Genome wide association studies elucidate links between genotypes and phenotypes. Recent studies point out the interest of conducting such experiments using k-mers as the base signal instead of single-nucleotide polymorphisms. kmdiff is a command line tool allowing efficient differential k-mer analyses on large sequencing cohorts.

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

Publication: hal-03885124

Contact: Pierre Peterlongo

Participants: Teo Lemane, Rayan Chikhi, Pierre Peterlongo

7.1.4 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

7.1.5 cdbgtricks

Keywords: De Bruijn graphs, Short reads

Functional Description: Cdbgtricks is a C++11 modular tool, mainly designed to update a compacted de Bruijn graph, when adding new sequences. In addition it indexes the graph, updates the index when adding sequences and so it enables querying sequences on the graph.

URL: https://github.com/khodor14/Cdbgtricks

Publication: hal-04765742

Contact: Pierre Peterlongo

7.1.6 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.

Contact: Pierre Peterlongo

7.1.7 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

7.1.8 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

7.1.9 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

7.1.10 gfagraphs

Keywords: Pangenomics, Variation graphs

Functional Description: This library aims to be an abstraction layer for the GFA file format, which is the standard file format for pangenome and variation graphs. It allows to load, save, modify and annotate a GFA file. Written in Python, it's goal is to provide an easy-to-use Graph object on which many operations can be performed.

Release Contributions: https://github.com/dubssieg/gfagraphs/commits/v0.3.0

URL: https://github.com/Tharos-ux/gfagraphs

Contact: Siegfried Dubois

Participants: Siegfried Dubois, Claire Lemaitre

Partner: INRAE

7.1.11 pancat

Name: PANgenome Comparison and Analysis Toolkit

Keywords: Pangenomics, Variation graphs

Functional Description: PANCAT is a command-line tool which allows to go through, visualize and compare pangenome graphs. Pangenome graphs (or variation graphs) are sequence graphs, encoded in a textual format, which describe shared and unique parts between a set of genomes. The aim of this tool is to answer technical and biological questions on this particular data structure.

Release Contributions: Changelog available here: https://github.com/dubssieg/pancat/compare/v0.2.0...v0.3.0

Contact: Siegfried Dubois

Participants: Siegfried Dubois, Claire Lemaitre

Partner: INRAE

7.1.12 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.

Contact: Siegfried Dubois

Partner: INRAE

7.1.13 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

7.1.14 HairSplitter

Keywords: Bioinformatics, Genome assembly, Bacterial strains, Metagenomics

Functional Description: HairSplitter takes as input a strain-oblivious assembly and sequencing reads and outputs a strain-separated assembly.

URL: https://github.com/RolandFaure/Hairsplitter

Contact: Roland Faure

7.1.15 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

7.1.16 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

8 New results

8.1 Data structures and indexing algorithms

8.1.1 FMSI k-mer indexing via masked superstrings

Participants: Karel Břinda.

The exponential growth of DNA sequencing data limits the searchable proportion of the data. In this context, tokenization of genomic data via their k-merization provides a path towards efficient algorithms for their compression and search. However, indexing even single k-mer sets still remains a significant bioinformatics challenge, especially if k-mer sets are sketched or subsampled. Here, we develop the FMSI index, a space-efficient data structure for unconstrained k-mer sets, based on approximated shortest superstrings and the Masked Burrows Wheeler Transform (MBWT), an adaptation of the BWT for masked superstrings. We implement this in a program called FMSI, and via extensive evaluations using prokaryotic pan-genomes, we show FMSI substantially improves space efficiency compared to the state of the art, while maintaining a competitive query time. Overall, our work demonstrates that superstring indexing is a highly general, parameter-free approach for modern k-mer sets, without imposing any constraints on their structure [47, 40].

8.1.2 Dynamic k-mer set operations via masked superstrings

Participants: Karel Břinda.

The design of efficient dynamic data structures for large k-mer sets belongs to central challenges of sequence bioinformatics. Recent advances in compact k-mer set representations via simplitigs/Spectrum-Preserving String Sets, culminating with the masked superstring framework, have provided data structures of remarkable space efficiency for wide ranges of k-mer sets. However, the possibility to perform set operations remained limited due to the static nature of the underlying compact representations. Here, we develop f-masked superstrings, a concept combining masked superstrings with custom demasking functions f to enable efficient k-mer set operations via string concatenation. Combined with the FMSI index for masked superstrings, we obtain a memory-efficient k-mer index supporting set operations via Burrows-Wheeler Transform merging. The framework provides a promising theoretical solution to a pressing bioinformatics problem and highlights the potential of f-masked superstrings to become an elementary data type for k-mer sets [47, 40, 48].

8.1.3 Optimized K-mer Matching For Million-Genome Collections On Laptops

Participants: Francesca Brunetti, Karel Břinda.

Bacteria are pivotal to human health, and their swift identification via genome sequencing technologies is critical, as evidenced by extensive bacterial genome databases like 661k and AllTheBacteria, which facilitate rapid diagnostics and epidemiological surveillance. However, real-time genome alignment on portable devices is hampered by existing technologies; while BLAST provides near-real-time searches, it requires substantial computational resources, and k-mer indexes, though advanced, fail to scale efficiently. Phylign offers a promising solution by managing alignments on a standard laptop efficiently, but its application is limited in critical point-of-care decisions due to dependencies like COBS, which are only optimal for short or near-exact matches. Our methodology utilizes phylogenetic compression to enhance k-mer matching across large genome collections on portable devices, aiming to significantly reduce processing times while retaining accuracy. We demonstrate through performance evaluations that replacing COBS with our phylogenetically compressed system can expedite k-mer matching by more than sixfold for extensive sequences, offering practical applications for biologists and epidemiologists using standard laptops [32].

8.1.4 Updating compacted de Bruijn graphs

Participants: Khodor Hannoush, Pierre Peterlongo.

We proposed Cdbgtricks, a new method for updating a compacted de Bruijn graph when adding novel sequences, such as full genomes. This method indexes the graph, enabling to identify in constant time the location (unitig and offset) of any queried k-mer. The update operation that is proposed also updates the index. Results show that Cdbgtricks is faster than main state-of-the-art tools, Bifrost and GGCAT. Cdbgtricks also benefits from the index of the graph to provide new functionalities, such as reporting the subgraph that shares a desired percentage of k-mers with a query sequence with the ability to query a set of reads. The open-source Cdbgtricks software is available at the following repository.

Cdbgtricks was presented during the Prague Stringology Conference and is published in the proceedings as well as in BiorXiv [27]. This work is also a main contribution of the PhD of Khodor Hannoush [43] and was also presented during seqBIM2024 [28].

8.1.5 The Backpack Quotient Filter: a dynamic and space-efficient data structure for querying k-mers with abundance

Participants: Victor Levallois, Pierre Peterlongo.

As databases storing genomic information, such as the European Nucleotide Archive, continue to grow exponentially, efficient solutions for data manipulation are imperative. One fundamental operation that remains challenging is querying these databases to determine the presence or absence of specific sequences and their abundance within datasets. We introduce a novel data structure indexing k-mers, the Backpack Quotient Filter (BQF), which serves as an alternative to the Counting Quotient Filter [57] (CQF). The BQF offers enhanced space efficiency compared to the CQF while retaining key properties, including abundance information and dynamicity, with a negligible false positive rate, below 10^{-5}%. The approach involves a redefinition of how abundance information is handled within the structure, alongside with an independent strategy for space efficiency. We show that the BQF uses 4x less space than the CQF on some of the most complex data to index: sea-water metagenomics sequences. Furthermore, we show that space efficiency increases as the amount of data to be indexed increases, which is in line with the original objective of scaling to ever-larger datasets [22].

8.1.6 Improve the resizing of the Backpack Quotient Filter

Participants: Nicolas Buchin, Victor Levallois, Pierre Peterlongo.

The BQF, presented section 8.1.5 is dynamic in nature, allowing for the addition of new items at any time. Upon reaching saturation of the structure, its capacity can be doubled. However, in the initial version of the BQF, this operation was not optimized, offering opportunities for improvements. In the work [33] presented at seqBIM2024, we proposed a novel algorithm that optimizes the resizing operation of the BQF data structure. The approach capitalizes on precomputed metadata. This makes it possible to efficiently determine the location of previously inserted values in the new resized structure. This drastically limits the number of steps required during this resizing operation, thus limiting the computation time, and also limiting the memory impact. The results show that this new approach enables one to gain a factor 4× to 8× in term of computation time. Notably, these tests have shown that the proposed enhancements enable to gain a factor 10 in terms of memory usage. It should be noted that this algorithm is not limited to the BQF, but can also be adapted to the Rank and Select Quotient Filter [57] and its variations, including the Counting Quotient Filter.

8.1.7 Back to sequences: Find the origin of k-mers

Participants: Pierre Peterlongo.

A vast majority of bioinformatics tools dedicated to the treatment of raw sequencing data heavily use the concept of k-mers. This enables us to reduce the redundancy of data (and thus the memory pressure), to discard sequencing errors, and to dispose of objects of fixed size that can be easily manipulated and compared to each other. A drawback is that the link between each k-mer and the original set of sequences to which it belongs is lost. Given the volume of data considered in this context, recovering this association is costly. In the work published in JOSS [13], we presented "back to sequences", a simple tool designed to index a set of k-mers of interest and to stream a set of sequences, extracting those containing at least one of the indexed k-mer. In addition, the occurrence positions of k-mers in the sequences can be provided. Our results show that "back to sequence" streams $\approx$ 200 short reads per millisecond, allowing us to search k-mers in hundreds of millions of reads in a matter of a few minutes.

8.1.8 Indexing and real-time user-friendly queries in terabyte-sized complex genomic datasets with kmindex and ORA.

Participants: Pierre Peterlongo.

Public sequencing databases contain vast amounts of biological information, yet they are largely underutilized as it is challenging to efficiently search them for any sequence(s) of interest. We presented kmindex, published in Nature Computation Science [21], an approach that can index thousands of metagenomes and perform sequence searches in a fraction of a second. The index construction is an order of magnitude faster than previous methods, while search times are two orders of magnitude faster. With negligible false positive rates below 0.01%, kmindex outperforms the precision of existing approaches by four orders of magnitude. In this work, we also demonstrated the scalability of kmindex by successfully indexing 1,393 marine seawater metagenome samples from the Tara Oceans project. Additionally, we introduce the publicly accessible web server Ocean Read Atlas, which enables real-time queries on the Tara Oceans dataset.

8.1.9 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, built using and running thanks to kmindex [21]. 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.

8.1.10 k-mer matrix compression

Participants: Alix Regnier, Pierre Peterlongo.

In the work presented at seqBIM2024, we propose a block compression method on top of the kmtricks matrices [21], 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 [52]. 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.

8.1.11 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 [56] 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 the work [31] presented during seqBIM2024, we discussed ongoing work on subquadratic compression of distance matrices of large bacterial genome collections. This approach takes advantage of the peculiar structure of those collections, that can extensively be explained by their underlying phylogeny. We 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. Overall, this work draws a path towards practical data structures that would be applicable to collections of millions of genomes with only negligible distance data distortion.

8.2 Sequence analysis algorithms

8.2.1 Haplotype assembly from long and noisy reads

Participants: Rumen Andonov, Roland Faure, Dominique Lavenier, Khac Minh Tam Truong.

Long-read assemblers face challenges in discerning closely related viral or bacterial strains, often collapsing similar strains into a single sequence. This limitation has been hampering metagenome analysis, as diverse strains may harbor crucial functional distinctions. We introduce a novel software, HairSplitter, designed to retrieve strains from a partially or totally collapsed assembly and long reads. The method uses a custom variant-calling process to operate with erroneous long reads and introduces a new read binning algorithm to recover an a priori unknown number of strains. On noisy long reads, HairSplitter recovers more strains while being faster than state-of-the-art tools, both in the cases of viruses and bacteria [17] [42]

In [41] we innovate by approaching strain separation as an Integer Linear Programming (ILP) problem. We introduce a strain-separation module, strainMiner, and integrate it into an established pipeline to create strain-separated assemblies from sequencing data. Across simulated and real experiments encompassing a wide range of error rates (5-12%), our tool consistently compared favorably to the state-of-the-art in terms of assembly quality and strain reconstruction. Moreover, strainMiner substantially cuts down the computational burden of strain-level assembly compared to published software by leveraging the powerful Gurobi solver.

8.2.2 Scaffolding step in genome assembly

Participants: Rumen Andonov, Victor Epain, Dominique Lavenier.

Scaffolding is an intermediate stage of fragment assembly. It consists in orienting and ordering the contigs obtained by the assembly of the sequencing reads. In the general case, the problem has been largely studied with the use of distance data between the contigs. Here we focus on a dedicated scaffolding for the chloroplast genomes. As these genomes are small, circular and with few repeats, numerous approaches have been proposed to assemble them. However, their specificities have not been sufficiently exploited. We give a new formulation for the scaffolding in the case of chloroplast genomes as a discrete optimisation problem, that we prove to be NP-Complete. It does not require distance information. It is focused on a genomic regions view, with the priority on scaffolding the repeats first. In this way, we encode the multimeric forms issue in order to retrieve several genome forms that can exist in the same chloroplast cell. In addition, we provide an Integer Linear Program (ILP) to obtain exact solutions that we implement in Python3 package khloraascaf. We test it on synthetic data to investigate its performance behaviour and its robustness against several chosen difficulties. While the scaffolding problem is traditionally defined with distance data, we show it is possible to avoid them in the case of the well-studied circular chloroplast genomes. The presented results show that the regions view seems to be sufficient to scaffold the repeats [16].

8.2.3 Assessing assembly quality in metagenomes of increasing complexity sequenced with HiFi long reads

Participants: Claire Lemaitre, Nicolas Maurice, Riccardo Vicedomini.

Evaluating the quality of metagenome assemblies can be a challenging task, especially when no reference genome is available and when comparing samples at various taxonomic complexity and sequencing depth. A high quality assembly is expected not only to produce high quality bins, but also

to be representative of most of the read sequences, especially in complex samples where algorithms struggle reconstructing low-abundance genomes. Although recent studies showed a great improvement in number and quality of bins obtained with novel highly accurate long reads (HiFi), it remains to be assessed how much of the sample these bins represent, especially in highly complex environmental samples.

We designed and implemented Mapler [53, 37], a metagenomic assembly and evaluation pipeline with a primary focus on evaluating the quality of HiFi-based metagenome assemblies. It incorporates state-of-the-art tools for assembly, binning, and assembly evaluation. In addition to classifying assembly bins in classical quality categories according to their marker gene content and taxonomic assignment, Mapler analyzes the alignment of reads on contigs. To do so, it calculates the ratio of mapped reads and bases, and separately analyzes mapped and unmapped reads via their k-mer frequency, read quality, and taxonomic assignment. We compared three metagenomic datasets of increasing complexity, all sequenced using PacBio HiFi technology: a mock community of 21 populations, a human gut microbiome, and a soil microbiome. At equal sequencing depth, for low and medium complexity samples (mock community and human gut, respectively), more than 90% of read sequences are found in bins regardless of the assembler used. In the soil sample, however, not only does the proportion of high quality bins drastically drop, but 54% to 88% of sequenced bases also fail to map to the assembly.

We show that considering multiple metrics is important for accurately assessing assembly quality, as relying on a single metric can be misleading. In particular, taking into account both bin quality and read alignment is indispensable, especially in complex environments. Mapler is a pipeline that calculates these metrics with minimal user input, producing both textual and graphical outputs, making it well-suited for evaluating assembly methodologies across a wide range of sample complexities. Mapler is open source and publicly available.

8.2.4 Construction of unitig matrices with abundance information

Participants: Riccardo Vicedomini.

Unitigs are biological sequences that compactly and exhaustively represent sequencing data or assembled genomes. They are constructed from k-mers but, unlike k-mers, they avoid the redundancy problem of multiple overlapping sequences covering the same genomic locus. They have proven useful for analyzing genomic diversity across several sequencing datasets. 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, by merging overlapping k-mers that unambiguously belong to the same sequence. Moreover, they preserve variations between samples while reducing disk space and reducing the number of rows in comparison to k-mer matrices. We addressed the limitations of current software by developing MUSET, 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 presence-absence unitig matrices. We evaluated MUSET's performance using datasets derived from a 618-GB collection of ancient oral sequencing samples, producing a filtered unitig matrix that records abundances in less than 10 hours and 20 GB 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 has been presented at seqBIM2024 and Genome Informatics. A manuscript has been provisionally accepted in the *Oxford Bioinformatics* journal. A pre-print version has been made available in [49].

8.2.5 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 aim to characterize 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 [34, 46].

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

8.2.6 Pangenome graph manipulation for local visualisation with pancat

Participants: Siegfried Dubois, Claire Lemaitre.

A pangenome graph is a data structure designed to represent variations among a collection of genomes, often from the same species. Pangenome graphs are enormous, with potentially billions of nodes and intricate connectivity patterns. Visualizing these data structures is a significant challenge, and needs to be addressed to facilitate pangenome graph analysis. Multiple visualization tools are currently available, each adopting different approaches: for example, some take paths into account, others focus on global topology, and some are based upon the representation of the embedded sequences. However, in many applications displaying the whole graph structure is not required and can be confusing. In this work, we propose several approaches to visualize local regions of a graph, along with a visualization tool. We introduce algorithms to aid in the local representation of pangenomes, compatible with state-of-the-art GFA visualizers. Our methods include subgraph extraction from positions in a genome, compression of small polymorphisms without loss inside a graph, and abstraction of parts of a graph to reduce the number of elements to render. Additionally, we propose a method allowing the simultaneous visualization of two pangenome graphs built from the same genomes. We tested these methods on real data, highlighting hotspots of differences between graphs made with different tools [35]. Implementations are part of a Python tool, pancat, available as a open-source project.

8.2.7 Understanding the limits of pangenome graphs for the analysis of large inversions

Participants: Fabrice Legeai, Claire Lemaitre, Sandra Romain.

Among structural variants, inversions are of particular interest in ecology studies as they can introduce phenotypic diversity and have a lasting impact on population genetics by locally impairing recombination. Pangenome graphs (PG) are a way to represent all scales of genomic diversity in a species. Several tools have been developed to construct PG from genome alignments and to detect variants from the graph topology. PG were shown to be particularly efficient for identifying and genotyping deletions and insertions in model organisms. However, they have not yet been thoroughly assessed on inversion polymorphism. We propose here to evaluate the ability of PG tools to represent and detect inversions in the case of a complex of butterfly species (*Coenonympha* genus) for which we already detected a dozen of large (> 100 kb) inversions. To compare the tools, we selected a chromosome with 2 large, 6 medium

sized (> 1 kb), and 9 small (< 1 kb) inversions, and built PG with 4 state of the art tools. Minigraph and minigraph-Cactus failed to build an accurate PG for such a degree of sequence divergence, while PGGB and Cactus found at most two inversions. In order to understand how PG handle inversions, we simulated the 17 known inversions in several synthetic chromosomes with increasing levels of single nucleotide divergence. We found that in such simplified graphs, most large simulated inversions are well represented with most tools. However, for smaller inversions and when the sequence divergence is higher, there is a significant variability in how the inversions are represented between PG tools. We analyzed the various obtained topological motifs, leading to methodological avenues for improving the detection of inversions in PG [39].

8.3 Information storage on DNA molecules

8.3.1 Design of long DNA molecules

Participants: Olivier Boulle, Dominique Lavenier, Julien Leblanc, Jacques Nicolas, Emeline Roux.

Generating long double-stranded DNA molecules with predefined sequences without a DNA template is particularly challenging. The DNA synthesis step often poses a bottleneck for various applications, including DNA-based data storage. Here, we present a fully in vitro protocol for synthesizing very long double-stranded DNA molecules, starting from commercially available short DNA blocks and completing the process in under three days using Golden Gate assembly. This novel approach enabled the efficient production of a 24 kb-long DNA molecule encoding a portion of the Declaration of the Rights of Man and of the Citizen from 1789. The resulting DNA molecule can be seamlessly cloned into an appropriate host vector system for amplification and selection [19]

8.3.2 Automation of biotechnology protocols

Participants: Olivier Boulle, Dominique Lavenier, Julien Leblanc, Jacques Nicolas.

Biotechnology experiments are often lengthy and labor-intensive. When performed manually by humans, they are prone to errors, and the extended duration of such experiments hinders scalability. To address these challenges, we have developed an automated platform, called DNAmaker, that streamlines the process. The system features a pipetting robot, a robotic arm that moves along a sliding rail, and a refrigerated storage area. All components are controlled by a Raspberry Pi system and can be programmed using Python language, offering both precision and flexibility.

Video of the DNAmaker platform

8.3.3 Random access with optimized primers

Participants: Dominique Lavenier.

Accessing specific DNA molecules from a large pool is an extremely complex task. This process relies on primers - short DNA sequences that serve as unique identifiers. Efficient random access requires high-quality primers, which are limited by the structural constraints of DNA. We have developed a method for designing primers that meet stringent biochemical standards, avoiding sequences likely to form undesirable secondary structures. The proposed tool relies on computer models to predict primer binding affinity and specificity, enabling users to adjust parameters according to laboratory protocols. This approach improves data retrieval efficiency and optimizes experimental workflows [36]

8.3.4 DNA caching

Participants: Olivier Boulle, Dominique Lavenier, Julien Leblanc, Jacques Nicolas.

The concept of DNA caching aims to improve the latency of DNA-based storage, addressing one of its primary limitations. Latency is the main drawback of DNA storage, but by drawing inspiration from caching techniques in computer science, it is possible to significantly accelerate access times. Not all files are accessed equally—some are requested far more frequently than others. Prioritizing the reduction of latency for these high-demand files is therefore essential. This approach is particularly feasible with nanopore sequencing: by storing frequently accessed files at higher concentrations within the DNA database, their retrieval times can be shortened, as nanopore sequencing operates in real time.

8.4 Processing-in-Memory

Participants: Charly Airault, Charles Deltel, Florestan de Moor, Erwan Drezen, Dominique Lavenier, Ulysse Mc Connell, Meven Mognol.

Processing-in-Memory Processing (MiP) consists of processing capabilities that are closely tied to the main memory. Unlike bringing all data into a centralized processor, which is far removed from data storage, in-memory computing processes data directly where it resides, eliminating most data movement, and thus dramatically improving the performance of massive data applications. NGS data analysis falls within these application areas, where PiM architecture can greatly accelerate key time-consuming software in the genomic and metagenomic fields. This year we explored the parallelization of several algorithms on a PiM server from the UPMEM company equipped with 20 UPMEM PiMs. These devices enhance standard 16GB DIMMs with 128 integrated computing units, leading to 160 GB of additional DRAM memory.

8.4.1 Sorting

Many genomic algorithms rely on k-mer counting and, more broadly, on sorting algorithms. In this work, we explored the parallelization of several sorting algorithms, including Quick Sort, Heap Sort, and Radix Sort, on the UPMEM Processing-in-Memory (PiM) architecture. Our findings highlight the potential benefits and limitations of sorting on PiM. UPMEM PiM can significantly enhance applications that are bound by speculative execution, making it well-suited for integrating sorting tasks into larger, more complex DPU-based applications without incurring performance losses. However, PiM sorting is not ideal as a standalone solution for accelerating sorting tasks.

8.4.2 Alignment of long reads

Sequence alignment is a fundamental task in genomic analysis, often constrained by the memory-wall in processor-centric architectures. We have developed a PiM-optimized version of the Needleman-Wunsch (N&W) algorithm tailored for long-read alignment, leveraging PiM devices developed by UPMEM. On the UPMEM server, This N&W implementation achieves performance improvements of an order of magnitude compared to traditional multicore server-based alignment tools [29].

8.4.3 Genome compression

As Next-Generation Sequencing (NGS) technologies continue to improve in accuracy and become widely integrated into healthcare infrastructures, it is crucial to develop efficient reference-based compressors. We have developed a fast, parallel mapper for compressing short reads, enabling lossless reference-based compression of NGS datasets such as Illumina reads. The mapper employs a non-exhaustive mapping approach against a reference genome to accelerate this step. To further enhance speed and reduce the number of sequence comparisons, the mapper integrates a Bloom filter-based dispatcher, which predicts

the genome regions most likely to match each read. Using real whole human sequencing datasets, we demonstrate that this approach achieves a speed-up of 1.2x to 2.7x compared to Genozip, the current leading state-of-the-art compressor, while maintaining a comparable compression ratio and reducing overall energy consumption [30]

8.5 Applications and bioinformatics analyses

8.5.1 Unlocking the Soil Microbiome: Unraveling Soil Microbial Complexity Using Long-Read Metagenomics.

Participants: Claire Lemaitre, Nicolas Maurice, Riccardo Vicedomini.

The soil microbiome remains poorly understood, but unraveling its genetic diversity is essential, given the pivotal functions primarily mediated through their protein arsenal. Although short-read (SR) shotgun metagenomics provided interesting insights into microbiome gene diversity, it fell short in delivering comprehensive microbial genome reconstructions. Metagenome-assembled genomes (MAGs) obtained from SR often yield fragmented assemblies and incomplete gene sets (over 90% of contigs smaller than 1kb and thus unusable for gene prediction). Using PacBio HiFi sequencing, we obtained long-reads (N50: 6kb) from tunnel culture soil metagenomes, surpassing the contig length of publicly available short-read metagenomes (N50: 1kb). We benchmarked three long-read assemblers for HiFi reads, including the recent MetaMDBG software and compared the results. Even if a substantial part of the reads remain unassembled, we succeeded in reconstructing dozens of (MAGs), which further enhanced reads contiguity encompassing bacterial, archaeal, and viral genomes from terrestrial environments. HiFi LR sequencing exhibits significant potential in elucidating the complexity of bacterial genome reconstruction. However, several critical considerations remain to be addressed such as the comprehensive scope of biodiversity captured by the LR approach compared to deep sequencing with SR [51, 50].

8.5.2 Chromosome-Level Assembly and Annotation of the Pearly Heath Coenonympha arcania Butterfly Genome

Participants: Fabrice Legeai, Claire Lemaitre, Sandra Romain.

We present the first chromosome-level genome assembly and annotation of a representative of the *Coenonympha* genus, the pearly heath *Coenonympha arcania* [20]. It is a relatively common and locally abundant species widely distributed in semi-open dry grasslands in Europe. Its genome was generated with a long-read PacBio HiFi sequencing approach, complemented with Hi-C scaffolding. We performed additional analyses on gene and repeat contents in comparison with 2 other high-quality *Satyrinae* genomes (*Maniola jurtina* and *Pararge aegeria*) and another Coenonympha species, the chestnut heath *C. glycerion*, all available from the Darwin Tree of Life project. Our genome comparative analysis revealed a high degree of chromosomal synteny between these genomes, suggesting very few large-scale chromosomal rearrangements between these 4 Satyrinae taxa.

This reference genome will enable future population genomics studies with Coenonympha butterflies, a species-rich genus that encompasses some of the most highly endangered butterfly taxa in Europe

8.5.3 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 analyzed the complete genome of the endosymbiotic bacteria *Spiroplasma ixodetis* that induces pea aphid (*Acyrthosiphon pisum*) male hosts killing during development. We performed comparative genomics analyses of *Spiroplasma* strains from *A. pisum* biotypes adapted to different host plants in order to reveal their phylogenetic associations and the diversity of putative virulence factors [12]. We also helped to process transcriptomics data in order to reveal four target genes that responded differently to photoperiod in two pea aphid lineages and putatively involved in the transitions from sexual to asexual reproduction [18]. We generated a chromosome-scale genome assembly for *Hyposoter didymator* that harbors a domesticated endogenous ichnovirus (HdIV), and showed that the virion is processed from 67 loci that are specifically amplified in the calyx cells during the wasp pupal stage. We also showed that this amplification required an ancestral virus gene [23]. We also produced the annotations of protein coding genes of a few Lepidoptera species and inspected the nuclear colinearity (synteny) of their genomes [25]. We characterized the genetic variations among a large population of virulent and avirulent strains of the oomycete *Plasmopara viticola*, which is responsible for the grapevine downy mildew. This work helped to identify a locus associated with resistant-breaking phenotypes [24]. We helped to identify important genes of oilseed rape (*Brassica napus*) that are related to drought stress during the seed maturity with regards to its infection by the telluric pathogen *Plasmodiophora brassicae* [14]. Finally, in strong collaboration with the Dyliss team, we developed and tested a data model able to store and query a large panel of data from genomics, transcriptomics or epigenomics experiments [26].

8.5.4 Rapid diagnostics of antibiotic resistance

Participants: Karel Břinda.

Timely diagnostic tools are needed to improve antibiotic treatment. Pairing metagenomic sequencing with genomic neighbor typing algorithms may support rapid clinically actionable results. We created resistance-associated sequence elements (RASE) databases for *Escherichia coli* and *Klebsiella spp.* and used them to predict antibiotic susceptibility in directly sequenced (Oxford Nanopore) urine specimens from critically ill patients. RASE analysis was performed on pathogen-specific reads from metagenomic sequencing. We evaluated the ability to predict (i) multi-locus sequence type (MLST) and (ii) susceptibility profiles. We used neighbor typing to predict MLST and susceptibility phenotype of *E. coli* (64/80) and *Klebsiella spp.* (16/80) from urine samples. When optimized by lineage score, MLST predictions were concordant for 73% of samples. Similarly, a RASE-susceptible prediction for a given isolate was associated with a specificity and a positive likelihood ratio (LR+) for susceptibility of 0.65 (95% CI, 0.54–0.76) and 2.26 (95% CI, 1.75–2.92), respectively, with an increase in the probability of susceptibility of 10%. A RASE-non-susceptible prediction was associated with a sensitivity and a negative likelihood ratio (LR-) for susceptibility of 0.79 (95% CI, 0.74–0.84) and 0.32 (95% CI, 0.24–0.43) respectively, with a decrease in the probability of susceptibility of 20%. Numerous antibiotic classes could reasonably be reconsidered empiric therapy by shifting empiric probabilities of susceptibility across relevant treatment thresholds. Moreover, these predictions can be available within 6 h. Metagenomic sequencing of urine specimens with neighbor typing provides rapid and informative predictions of lineage and antibiotic susceptibility with the potential to impact clinical decision-making [15].

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 research visitors

10.1.1 Visits of international scientists

Other international visits to the team

Arya Kaul

Status PhD student

Institution of origin: Harvard Medical School

Country: USA

Dates: Sep 2023-Feb 2024

Context of the visit: visiting PhD student with K. Břinda

Mobility program/type of mobility: Chateaubriand fellowship

Francesca Brunetti

Status PhD student

Institution of origin: Sapienza University of Rome

Country: Italy

Dates: Nov 2023-Nov 2024

Context of the visit: visiting PhD student with K. Břinda

Mobility program/type of mobility: Sapienza PhD mobility

Veronika Hendrychová

Status MSc student

Institution of origin: Czech Technical University in Prague

Country: Czech Republic

Dates: Sep 2023-June 2024

Context of the visit: visiting MSc. student with K. Břinda

Mobility program/type of mobility: Erasmus

Ondřej Sladký

Status BSc student

Institution of origin: Charles University

Country: Czech Republic

Dates: Feb 19-21, 2024

Context of the visit: short collaborative visit with K. Břinda

Mobility program/type of mobility: scientific visit

Josipa Lipovac

Status PhD student

Institution of origin: University of Zagreb

Country: Croatia

Dates: Nov 2024-April 2025

Context of the visit: visiting PhD student with R. Vicedomini

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

10.2 European initiatives

10.2.1 H2020 projects

ALPACA ALPACA project on cordis.europa.eu

Title: ALgorithms for PAngenome Computational Analysis

Duration: From January 1, 2021 to December 31, 2024

Partners:
- INSTITUT NATIONAL DE RECHERCHE EN INFORMATIQUE ET AUTOMATIQUE (INRIA), France
- HEINRICH-HEINE-UNIVERSITAET DUESSELDORF (UDUS), Germany
- HELSINGIN YLIOPISTO, Finland
- THE CHANCELLOR MASTERS AND SCHOLARS OF THE UNIVERSITY OF CAMBRIDGE, United Kingdom
- EUROPEAN MOLECULAR BIOLOGY LABORATORY (EMBL), Germany
- GENETON S.R.O. (Geneton), Slovakia
- UNIVERSITA DI PISA (UNIPI), Italy
- UNIVERZITA KOMENSKEHO V BRATISLAVE (UK BA), Slovakia
- INSTITUT PASTEUR (IP), France
- UNIVERSITA' DEGLI STUDI DI MILANO-BICOCCA (UNIMIB), Italy
- CENTRE NATIONAL DE LA RECHERCHE SCIENTIFIQUE CNRS (CNRS), France
- UNIVERSITAET BIELEFELD (UNIBI), Germany
- STICHTING NEDERLANDSE WETENSCHAPPELIJK ONDERZOEK INSTITUTEN (NWO-I), Netherlands

Inria contact: Pierre Peterlongo

Coordinator: Alexander Schönhuth

Summary: Genomes are strings over the letters A,C,G,T, which represent nucleotides, the building blocks of DNA. In view of ultra-large amounts of genome sequence data emerging from ever more and technologically rapidly advancing genome sequencing devices—in the meantime, amounts of sequencing data accrued are reaching into the exabyte scale—the driving, urgent question is: how can we arrange and analyze these data masses in a formally rigorous, computationally efficient and biomedically rewarding manner?

Graph based data structures have been pointed out to have disruptive benefits over traditional sequence based structures when representing pan-genomes, sufficiently large, evolutionarily coherent collections of genomes. This idea has its immediate justification in the laws of genetics: evolutionarily closely related genomes vary only in relatively little amounts of letters, while sharing the majority of their sequence content. Graph-based pan-genome representations that allow to

remove redundancies without having to discard individual differences, make utmost sense. In this project, we will put this shift of paradigms—from sequence to graph based representations of genomes—into full effect. As a result, we can expect a wealth of practically relevant advantages, among which arrangement, analysis, compression, integration and exploitation of genome data are the most fundamental points. In addition, we will also open up a significant source of inspiration for computer science itself.

For realizing our goals, our network will (i) decisively strengthen and form new ties in the emerging community of computational pan-genomics, (ii) perform research on all relevant frontiers, aiming at significant computational advances at the level of important breakthroughs, and (iii) boost relevant knowledge exchange between academia and industry. Last but not least, in doing so, we will train a new, "paradigm-shift-aware" generation of computational genomics researchers.

BioPIM link

Title: Processing-in-memory architectures and programming libraries for bioinformatics algorithms

Duration: From May 1, 2022 to April 30, 2025

Partners:

- Bilkent University
- ETH Zürich
- Pasteur Institute
- CNRS
- IBM Research Zürich
- Technion - Israel Institute of Technology
- UPMEM company

Inria contact: Dominique Lavenier

Coordinator: Can Alkan (Bilkent University)

Summary: The BioPIM project aims to leverage the emerging processing-in-memory (PIM) technologies to enable powerful edge computing. The project will focus on co-designing algorithms and data structures commonly used in bioinformatics together with several types of PIM architectures to obtain the highest benefit in cost, energy, and time savings. BioPIM will also impact other fields that employ similar algorithms. Designs and algorithms developed during the BioPIM project will not be limited to chip hardware: they will also impact computation efficiency on all forms of computing environments including cloud platforms.

10.3 National initiatives

10.3.1 PEPR

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

Participants: Olivier Boullé, Dominique Lavenier, Julien Leblanc, Jacques Nicolas, Emeline Roux.

Coordinators: Marc Antonini

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

Partners: I3S, LabSTIC, IMT-Atlantique, GenScale 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) et 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).

10.3.2 ANR

Project SeqDigger: Search engine for genomic sequencing data

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

Coordinator: P. 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 Divalps: diversification and adaptation of alpine butterflies along environmental gradients

Participants: Fabrice Legeai, Claire Lemaitre, Sandra Romain.

Coordinator: L. Després (Laboratoire d'écologie alpine (LECA), UMR CNRS 5553, Grenoble)

Duration: 42 months (Jan. 2021 – Dec. 2024)

Partners: LECA, UMR CNRS 5553, Grenoble; CEFE, UMR CNRS 5175, Montpellier; Genscale Inria/IRISA Rennes.

Description: The Divalps project aims at better understanding how populations adapt to changes in their environment, and in particular climatic and biotic changes with altitude. Here, we focus on a complex of butterfly species distributed along the alpine altitudinal gradient. We will analyse the genomes of butterflies in contact zones to identify introgressions and rearrangements between taxa.

GenScale's task is to develop new efficient methods for detecting and representing the genomic diversity among this species complex. We will focus in particular on Structural Variants and genome graph representations.

Project GenoPIM: Processing-in-Memory for Genomics

Participants: Charly Airault, Charles Deltel, 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 Břinda

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.

10.3.3 Inria Exploratory Action

Défi Inria OmicFinder

Participants: Pierre Peterlongo, Victor Levallois, 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 Břinda

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.

10.4 Regional initiatives

Labex CominLabs: dnarXiv project

Participants: Guillermo Barroso, Olivier Boullé, Dominique Lavenier, Julien Leblanc, Jacques Nicolas.

Coordinator: Dominique Lavenier

Duration: 60 months (2021-2026)

Description: The dnarXiv project aims at exploring data storage on DNA molecules. This kind of storage has the potential to become a major archive solution in the mid-to long term. In this project, two key promising biotechnologies are considered: enzymatic DNA synthesis and DNA nanopore sequencing. The objective is to propose advanced solutions in terms of coding schemes (i.e., source and channel coding) and data security (i.e., data confidentiality/integrity and DNA storage authenticity), that consider the constraints and advantages of the chemical processes and biotechnologies involved in DNA storage.

Rennes Metropole: Allocation d'installation scientifique

Participants: Karel Břinda.

Coordinator: Karel Břinda

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.

Défi scientifique RECHERCHE TRANSDISCIPLINAIRE INTERPÔLES, Université de Rennes

Participants: Emeline Roux.

Coordinator: Emeline Roux

Duration: 24 months (January 2023 - December 2024)

Partners: Inria teams: Dyliss.

External partner is Institut NuMeCan.

Description: The role of the microbiota, particularly the intestinal microbiota, in the physiology and pathophysiology of numerous diseases has been recognized for over 15 years now. The transdisciplinary research project initiated the development of a pipeline for the analysis of long-read sequencing data from intestinal microbiota. The aim is to determine the precise meta-genome (at strain level) and consequently predict the meta-metabolome of the intestinal microbiota. The method developed will provide a better understanding of food-microbiota-host interactions, with a view to proposing innovative solutions in the future for restoring an intestinal microbiota with complete metabolic functions.

11 Dissemination

11.1 Promoting scientific activities

Participants: Karel Břinda, Siegfried Dubois, Khodor Hannoush, Dominique Lavenier, Fabrice Legeai, Claire Lemaitre, Victor Levallois, Nicolas Maurice, Jaques Nicolas, Pierre Peterlongo, Riccardo Vicedomini.

11.1.1 Scientific events: organisation

General chair, scientific chair

- seqBIM2024: national meeting of the sequence algorithms GT seqBIM, Rennes, Nov 2024 (90 participants, 2 days) [C. Lemaitre]

Member of the organizing committees

- seqBIM2024: national meeting of the sequence algorithms GT seqBIM, Rennes, Nov 2024 (90 participants, 2 days) [K. Břinda, S. Dubois, C. Lemaitre, V. Levallois, P. Peterlongo, R. Vicedomini]

11.1.2 Scientific events: selection

Chair of conference program committees

- seqBIM2024: national meeting of the sequence algorithms GT seqBIM, Rennes, Nov 2024 (90 participants, 2 days) [K. Břinda, C. Lemaitre, P. Peterlongo, R. Vicedomini]

Member of the conference program committees

- seqBIM2024: national meeting of the sequence algorithms GT seqBIM, Rennes, Nov 2024 (90 participants, 2 days) [K. Břinda, S. Dubois, C. Lemaitre, V. Levallois, P. Peterlongo, R. Vicedomini]
- ISMB 2024: 32th Annual Conference on Intelligent Systems for Molecular Biology, Montréal, Canada, 2024 [C. Lemaitre, P. Peterlongo]
- RECOMB-Seq 2024: 14th RECOMB Satellite Conference on Biological Sequence Analysis, UK, 2024 [C. Lemaitre]

Reviewer

- WABI 2024 [K. Hannoush, R. Vicedomini]

11.1.3 Journal

Reviewer - reviewing activities

- Bioinformatics [D. Lavenier]
- BMC Bioinformatics [D. Lavenier]
- Drug Discovery Today [D. Lavenier]
- Genome Biology [C. Lemaitre]
- Journal of Supercomputing [D. Lavenier]
- Journal of Open Source Software [P. Peterlongo]
- Nature Communications [C. Lemaitre]
- PCI Math Comp Biol [K. Břinda, R. Vicedomini]
- PLoS Computational Biology [K. Břinda]

11.1.4 Invited talks

- D. Lavenier, "DNA Storage", CORESA 2024, COmpresseion et REpresentation des Signaux Audiovisuels, Rennes, Nov. 2024
- D. Lavenier, "Information Storage on DNA", Journée scientifique du L2TI, Paris, Dec. 2024
- P. Peterlongo, "Index the planet: index SRA unitigs with kmindex", Genome Informatics, Hixton UK, Nov. 2024
- P. Peterlongo, "Indexing and real-time user-friendly queries in terabyte-sized complex genomic datasets", Labri Seminar, Bordeaux, May 2024

11.1.5 Leadership within the scientific community

- Members of the Scientific Advisory Board of the GDR BIMMM (National Research Group in Molecular Bioinformatics) [P. Peterlongo, C. 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) [C. Lemaitre]
- Animator of the INRAE Center for Computerized Information Treatment "BARIC" [F. Legeai]
- Member of the PEPR MoleculArxiv Executive Committee [D. Lavenier]

11.1.6 Scientific expertise

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

11.1.7 Research administration

- Corresponding member of COERLE (Inria Operational Committee for the assessment of Legal and Ethical risks) [J. Nicolas]
- Member of the steering committee of the INRAE BIPAA Platform (BioInformatics Platform for Agro-ecosystems Arthropods) [P. Peterlongo]
- Scientific Advisor of The GenOuest Platform (Bioinformatics Resource Center of BioGenOuest) [P. Peterlongo]
- Chair of the committee in charge of all the temporary recruitments (“Commission Personnel”) at Inria Rennes-Bretagne Atlantique and IRISA [D. Lavenier]

11.2 Teaching - Supervision - Juries

Participants: Rumen Andonov, Karel Břinda, Siegfried Dubois, Dominique Lavenier, Fabrice Legeai, Claire Lemaitre, Victor Levallois, Nicolas Maurice, Jaques Nicolas, Pierre Peterlongo, Emeline Roux, Riccardo Vicedomini.

11.2.1 Teaching administration

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

11.2.2 Teaching

- Licence: R. Andonov, Linear Programming, 30h, L3 Miage, Univ. Rennes, France.
- Licence : E. Roux, Biochemistry, 90h, L1 and L3, Univ. Rennes, France.
- Licence: N. Maurice, Object Oriented Programming, 38h, L2, INSA Rennes, France.
- Licence: S. Dubois, Introduction to Java, 24h, L1 Informatics, Univ. Rennes, France.
- Master: S. Dubois, Object-Oriented Programming, 32h, M2 Bioinformatics, Univ. Rennes, France.
- Master: J. Nicolas, Logic, ASP and CSP, 36h, M1 Artificial Intelligence, Univ. Rennes, France.
- Master: C. Lemaitre, P. Peterlongo, V. Levallois, Sequence Bioinformatics, 42h, M1 Informatics, Univ. Rennes, France.
- Master: C. Lemaitre, P. Peterlongo, Algorithms on Sequences, 52h, M2 Bioinformatics, Univ. Rennes, France.
- Master : R. Vicedomini, K. Břinda, Experimental Bioinformatics, 21h, M1, ENS Rennes, France.
- Master : E. Roux, biochemistry and microbiology, 130h, M1 and M2, Univ. Rennes, France.

11.2.3 HDR defense

- HDR: E. Roux, Food and Functionalities [45], defended:29/11/2024.

11.2.4 Supervision

- PhD: S. Romain, Identification, genotyping and representation of structural variants in pangenomes [44], Inria (ANR Divalps), defended: 08/07/2024, C. Lemaitre, F. Legeai.
- PhD: K. Hannoush, Dynamic pan-genome graphs [43], Inria (ITN Alpaca), defended: 13/12/20204, P. Peterlongo, C. Marchet.
- PhD: R. Faure, Haplotype assembly from long reads [42], Univ Rennes, defended: 27/11/2024, D. Lavenier, J-F. Flot.
- PhD: C. Duitama (Sequence Bioinformatics group, Institut Pasteur, Paris), Algorithms based on k-mers for ancient oral metagenomics : Tools for contamination removal and assessment in palaeometagenomics [duitamagonzalez:tel-04560480v2], defended: 30/01/2024, R. Chikhi, H. Richard, R. Vicedomini.
- PhD in progress: M. Mognol, Processing-in-Memory, CIFRE, since 01/04/2022, D. Lavenier.
- PhD in progress: S. Dubois, Characterizing structural variation in pangenome graphs, Inria (PEPR-ANR Agrodiv), since 15/09/2023, C. Lemaitre, T. Faraut, M. Zynticki.
- PhD in progress: N. Maurice, Sequence algorithmics for de novo genome assembly from complex metagenomic data, Inria (PEPR-ANR Mistic), since 01/10/2023, C. Lemaitre, C. Frioux, R. Vicedomini.
- PhD in progress: V. Levallois, Indexing genomic data, Défi Inria OmicFinder, since 01/10/2023, P. 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, F. Legeai.
- PhD in progress: A. Regnier, Limiting the size of data structures indexing genomic sequences, Inria (PEPR-ANR Agrodiv), since 01/10/2024, P. Peterlongo.

- PhD in progress: L. Ackermann, Developing efficient algorithms for sublinear search in large genome databases, Inria, since 01/10/2024, K. Břinda, P. Peterlongo.
- PhD in progress: T. Truong, Computational methods for phylogenetic compression, Univ Rennes, since 01/11/2024, K. Břinda, P. Peterlongo, D. Lavenier.
- PhD in progress: L. Angevin, Fine characterization of the intestinal microbiota and prediction of metabolite production, Univ Rennes, since 01/10/2024, E. Roux, R. Vicedomini, P. Peterlongo.
- PhD in progress: M. Temperville, Methods for characterizing structural variations in genomes with linked-read data, Univ Rennes, since since 01/10/2024, F. Legeai, C. Lemaitre, C. Mérot.

11.2.5 Juries

- *Reviewer of HDR thesis*
  - Sylvain Foissac, Univ. Toulouse, June 2024 [C. Lemaitre]
- *Member of HDR thesis jury*
  - Loïs Maignien, Univ. Brest, Dec. 2024 [C. Lemaitre]
- *President of PhD thesis jury*
  - C. Duitama, Sorbonne University, Jan 2024 [P. Peterlongo]
  - S. Romain, Univ. Rennes, Nov 2024 [P. Peterlongo]
  - K. Hannoush, Univ. Rennes, Dec 2024 [C. Lemaitre]
  - X. Pic, Univ. Nice, Sep 2024 [D. Lavenier]
- *Reviewer of PhD thesis*
  - C. Duitama, Sorbonne University, Jan 2024 [P. Peterlongo]
  - M. Leinonen, Univ. Helsinki (Finland), Feb 2024 [R. Vicedomini]
  - D. Martincigh, Univ. Udine (Italy), Sep 2024 [R. Vicedomini]
- *Member of PhD thesis jury*
  - Berton, IMT-A, Univ. Brest [D. Lavenier]
  - D. Martincigh, Univ. Udine (Italy), Sep 2024 [R. Vicedomini]
- *Member of PhD thesis committee*
  - T. Baudeau, Univ. Lille [C. Lemaitre]
  - Léa Nicolas, Univ Rennes [C. Lemaitre]
  - Dimple Adiwal, Univ Rennes [C. Lemaitre]
  - Jules Burgat, Univ Rennes [C. Lemaitre]
  - Florent Couturier, Univ Bordeaux [C. Lemaitre]
  - L. Vandame, Univ. Lille [P. Peterlongo]
  - O. Weppe, Univ. Rennes [P. Peterlongo]
  - A. Dequay, Univ. Rennes [D. Lavenier]
  - A. Ezzeddine, IMT-A, Univ. Brest [D. Lavenier]
  - H. Gasnier, IMT-A, Univ. Brest [D. Lavenier]
  - L. de la Fuenté, Univ. Rennes [D. Lavenier]
  - R. Jaafar, Univ. Rennes [R. Vicedomini]

11.3 Popularization

Participants: Dominique Lavenier, Pierre Peterlongo, Emeline Roux.

11.3.1 Specific official responsibilities in science outreach structures

- Member of the Interstice editorial board [P. Peterlongo]

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

- The dnarXiv project: Video [D. Lavenier]
- Article Interstices [54]: Indexer les données génomiques : un défi de taille à relever. [P. Peterlongo]
- Article Interstices [55]: Indexer des milliards d'éléments avec les filtres de Bloom. [P. Peterlongo]

11.3.3 Participation in Live events

- Chiche! Interventions in high school classes to make high school students aware of research careers in the digital sector. Two interventions made in 2024. [P. Peterlongo]
- Maison Pour La Science: Training for school teachers on sport, nutrition and health. Two interventions made in 2024. [E. Roux]
- Les Champs Libres, Espace des Sciences: Sugar, between pleasure and harm. One intervention made in 2024. [E. Roux]
- Programming sessions LCLC for Middle school female students [L. Ackermann]

12 Scientific production

12.1 Major publications

[1] R. Andonov, H. Djidjev, S. François and D. Lavenier. 'Complete Assembly of Circular and Chloroplast Genomes Based on Global Optimization'. In: *Journal of Bioinformatics and Computational Biology* (2019), pp. 1–28. DOI: 10.1142/S0219720019500148. URL: https://hal.archives-ouvertes.fr/hal-02151798 (cit. on p. 5).

[2] G. Benoit, C. Lemaitre, D. Lavenier, E. Drezen, T. Dayris, R. Uricaru and G. Rizk. 'Reference-free compression of high throughput sequencing data with a probabilistic de Bruijn graph'. In: *BMC Bioinformatics* 16.1 (Sept. 2015). DOI: 10.1186/s12859-015-0709-7. URL: https://hal.inria.fr/hal-01214682 (cit. on p. 5).

[3] 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 (Nov. 2016). DOI: 10.7717/peerj-cs.94. URL: https://hal.inria.fr/hal-01397150 (cit. on p. 5).

[4] R. Chikhi and G. Rizk. 'Space-efficient and exact de Bruijn graph representation based on a Bloom filter'. In: *Algorithms for Molecular Biology* 8.1 (2013), p. 22. DOI: 10.1186/1748-7188-8-22. URL: http://hal.inria.fr/hal-00868805 (cit. on pp. 4, 5).

[5] E. Drezen, G. Rizk, R. Chikhi, C. Deltel, C. Lemaitre, P. Peterlongo and D. Lavenier. 'GATB: Genome Assembly & Analysis Tool Box'. In: *Bioinformatics* 30 (2014), pp. 2959–2961. DOI: 10.1093/bioinformatics/btu406. URL: https://hal.archives-ouvertes.fr/hal-01088571 (cit. on pp. 4, 5).

[6] C. Guyomar, F. Legeai, E. Jousselin, C. 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 (Dec. 2018). DOI: 10.1186/s40168-018-0562-9. URL: https://hal.archives-ouvertes.fr/hal-01926402 (cit. on p. 6).

[7] D. Lavenier. 'DNA Storage: Synthesis and Sequencing Semiconductor Technologies'. In: IEDM 2022 - 68th Annual IEEE International Electron Devices Meeting. San Francisco, United States: IEEE, 3rd Dec. 2022, pp. 1–4. URL: https://hal.science/hal-03902786 (cit. on p. 5).

[8] 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 pp. 4, 5, 7).

[9] 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. London, United Kingdom, June 2017, pp. 1–11. URL: https://hal.inria.fr/hal-01566246 (cit. on p. 4).

[10] 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. URL: https://hal.inria.fr/hal-01081089 (cit. on p. 5).

[11] 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* (Nov. 2014), pp. 1–12. DOI: 10.1093/nar/gku1187. URL: https://hal.inria.fr/hal-01083715 (cit. on p. 5).

12.2 Publications of the year

International journals

[12] H. Arai, F. Legeai, D. Kageyama, A. Sugio and J.-C. Simon. 'Genomic insights into Spiroplasma endosymbionts that induce male-killing and protective phenotypes in the pea aphid'. In: *FEMS Microbiology Letters* 371 (9th Jan. 2024). DOI: 10.1093/femsle/fnae027. URL: https://inria.hal.science/hal-04818782 (cit. on p. 22).

[13] A. Baire, P. Marijon, F. Andreace and P. Peterlongo. 'Back to sequences: Find the origin of k-mers'. In: *Journal of Open Source Software* 9.101 (23rd Sept. 2024), p. 7066. DOI: 10.21105/joss.07066. URL: https://inria.hal.science/hal-04707743 (cit. on p. 14).

[14] G. Bianchetti, V. Clouet, F. Legeai, C. Baron, K. Gazengel, B. Ly Vu, S. Baud, A. To, M. J. Manzanares-Dauleux, J. Buitink and N. Nesi. 'Identification of transcriptional modules linked to the drought response of Brassica napus during seed development and their mitigation by early biotic stress'. In: *Physiologia Plantarum*. Physiologia Plantarum 176.1 (2024), e14130. DOI: 10.1111/ppl.14130. URL: https://hal.science/hal-04443649 (cit. on p. 22).

[15] A. C. Carroll, L. Mortimer, H. Ghosh, S. Reuter, H. Grundmann, K. Brinda, W. P. Hanage, A. Li, A. Paterson, A. Purssell, A. Rooney, N. R. Yee, B. Coburn, S. Able-Thomas, M. Antonio, A. Mcgeer and D. R. Macfadden. 'Rapid inference of antibiotic susceptibility phenotype of uropathogens using metagenomic sequencing with neighbor typing'. In: *Microbiology Spectrum* (29th Nov. 2024), pp. 1–16. DOI: 10.1128/spectrum.01366-24. URL: https://inria.hal.science/hal-04829427 (cit. on p. 22).

[16] V. Epain and R. Andonov. 'Global exact optimisations for chloroplast structural haplotype scaffolding'. In: *Algorithms for Molecular Biology* 19.5 (6th Feb. 2024), pp. 1–35. DOI: 10.1186/s13015-023-00243-1. URL: https://inria.hal.science/hal-04134429 (cit. on p. 16).

[17] R. Faure, D. Lavenier and J.-F. Flot. 'HairSplitter: haplotype assembly from long, noisy reads'. In: *Peer Community In Mathematical and Computational Biology* (14th Feb. 2024), pp. 1–16. DOI: 10.1101/2024.02.13.580067. URL: https://hal.science/hal-04739354 (cit. on p. 16).

[18] M. D. Huguet, S. Robin, S. Hudaverdian, S. Tanguy, N. Prunier-Leterme, R. Cloteau, S. Baulande, P. Legoix-Né, F. Legeai, J.-C. Simon, J. Jaquiéry, D. Tagu and G. Le Trionnaire. 'Transcriptomic basis of sex loss in the pea aphid'. In: *BMC Genomics*. Bmc Genomics 25.1 (21st Feb. 2024), p. 202. DOI: 10.1186/s12864-023-09776-6. URL: https://hal.science/hal-04484084 (cit. on p. 22).

[19] J. Leblanc, O. Boulle, E. Roux, J. Nicolas, D. Lavenier and Y. Audic. 'Fully in vitro iterative construction of a 24 kb-long artificial DNA sequence to store digital information'. In: *Biotechniques* 76.5 (2024), pp. 203–215. DOI: 10.2144/btn-2023-0109. URL: https://hal.science/hal-04567229 (cit. on p. 19).

[20] F. Legeai, S. Romain, T. Capblancq, P. Doniol-Valcroze, M. Joron, C. Lemaitre and L. Després. 'Chromosome-Level Assembly and Annotation of the Pearly Heath Coenonympha arcania Butterfly Genome'. In: *Genome Biology and Evolution* 16.3 (Mar. 2024), evae055. DOI: 10.1093/gbe/evae055. URL: https://hal.science/hal-04530573 (cit. on p. 21).

[21] 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 pp. 14, 15).

[22] V. Levallois, F. Andreace, B. Le Gal, Y. Dufresne and P. Peterlongo. 'The Backpack Quotient Filter: a dynamic and space-efficient data structure for querying k -mers with abundance'. In: *iScience* (20th Feb. 2024), p. 111435. DOI: 10.1016/j.isci.2024.111435. URL: https://pasteur.hal.science/pasteur-04698948 (cit. on p. 13).

[23] A. Lorenzi, F. Legeai, V. Jouan, P.-A. Girard, M. R. Strand, M. Ravallec, M. Eychenne, A. Bretaudeau, S. Robin, J. Rochefort, M. Villegas, G. R. Burke, R. Rebollo, N. Nègre and A.-N. Volkoff. 'Identification of a viral gene essential for the genome replication of a domesticated endogenous virus in ichneumonid parasitoid wasps'. In: *PLoS Pathogens* 20.4 (25th Apr. 2024), e1011980. DOI: 10.1371/journal.ppat.1011980. URL: https://hal.inrae.fr/hal-04572197 (cit. on p. 22).

[24] M. Paineau, A. Minio, P. Mestre, F. Fabre, I. D. Mazet, C. Couture, F. Legeai, T. Dumartinet, D. Cantu and F. Delmotte. 'Multiple deletions of candidate effector genes lead to the breakdown of partial grapevine resistance to downy mildew'. In: *New Phytologist* 243.4 (21st June 2024), pp. 1490–1505. DOI: 10.1111/nph.19861. URL: https://hal.inrae.fr/hal-04633322 (cit. on p. 22).

[25] C. Perrier, R. Allio, F. Legeai, M. Gautier, F. Bénéluz, W. Marande, A. Théron, N. Rodde, M. Herrera, L. Saune, 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* (2025), esae069. DOI: 10.1093/jhered/esae069. URL: https://cnrs.hal.science/hal-04792049. In press (cit. on p. 22).

[26] Y. Tirlet, M. Boudet, E. Becker, F. Legeai and O. Dameron. 'Generic and queryable data integration schema for transcriptomics and epigenomics studies'. In: *Computational and Structural Biotechnology Journal* 23 (Dec. 2024), pp. 4232–4241. DOI: 10.1016/j.csbj.2024.11.022. URL: https://inria.hal.science/hal-04818860 (cit. on p. 22).

International peer-reviewed conferences

[27] K. Hannoush, C. Marchet and P. Peterlongo. 'Cdbgtricks: Strategies to update a compacted de Bruijn graph'. In: *Prague Stringology Conference 2024*. PSC 2024 - Prague Stringology Conference. Prague (CZ), Czech Republic, 24th Aug. 2024, pp. 1–13. DOI: 10.1101/2024.05.24.595676. URL: https://hal.science/hal-04765742 (cit. on p. 13).

[28] K. Hannoush, C. Marchet and P. Peterlongo. 'Strategies to update a compacted de Bruijn graph'. In: *SeqBIM 2024*. SeqBIM 2024 - Journées annuelles du groupe de travail SeqBIM (Séquences en Bioinformatique, Informatique et Mathématiques). Rennes, France, 19th Nov. 2024, pp. 1–2. URL: https://inria.hal.science/hal-04861344 (cit. on p. 13).

[29] M. Mognol, D. Lavenier and J. Legriel. 'Parallelization of the Banded Needleman & Wunsch Algorithm on UPMEM PiM Architecture for Long DNA Sequence Alignment'. In: ICPP 2024 - International Conference on Parallel Processing. Gotland, Sweden: ACM, 2024, pp. 1–10. URL: https://hal.science/hal-04739328 (cit. on p. 20).

[30] F. de Moor, M. Mognol, C. Deltel, E. Drezen, J. Legriel and D. Lavenier. 'MiMyCS: A Processing-in-Memory Read Mapper for Compressing Next-Gen Sequencing Datasets'. In: *IEEEXplore*. BIBM 2024 - IEEE International Conference on Bioinformatics and Biomedicine. Vol. BIBM2024. 1. Lisbonne, Portugal: IEEE, 3rd Dec. 2024, p. 7176. URL: https://inria.hal.science/hal-04821180 (cit. on p. 21).

Conferences without proceedings

[31] L. Ackermann, P. Peterlongo and K. Břinda. 'Towards space-efficient data structures for large genome-distance matrices with quick retrieval'. In: SeqBim 2024 - Journées sur les Séquences en Bioinformatique, Informatique et. Mathématiques. Rennes, France, 2024. URL: https://hal.science/hal-04819782 (cit. on p. 15).

[32] F. Brunetti and K. Břinda. 'Optimized K-mer Matching For Million-Genome Collections On Laptops'. In: SeqBIM 2024 - Journées sur les Séquences en Bioinformatique, Informatique et. Mathématiques, Rennes, France, 2024, pp. 1–2. URL: https://inria.hal.science/hal-04842871 (cit. on p. 13).

[33] N. Buchin, V. Levallois and P. Peterlongo. 'Improve the resizing of the Backpack Quotient Filter'. In: SeqBim 2024 - Journées annuelles du groupe de travail SeqBIM (Séquences en Bioinformatique, Informatique et Mathématiques). Rennes, France, 2024. URL: https://hal.science/hal-04860174 (cit. on p. 14).

[34] S. Dubois, Z. Matthias, C. Lemaitre and F. Thomas. 'Towards an edit distance between pangenome graphs'. In: DSB 2024 - Workshop Data Structures in Bioinformatics. Montpellier, France, 13th Mar. 2024, pp. 1–1. URL: https://hal.science/hal-04725596 (cit. on p. 18).

[35] S. Dubois, M. Zytnicki, T. Faraut and C. Lemaitre. 'Pangenome graph manipulation for local visualisation with pancat'. In: MIGGS1 2024 - 1st symposium on Methods for Interfacing with Graphs of Genomic Sequences. Lille, France, 2024, pp. 1–1. URL: https://hal.science/hal-04724519 (cit. on p. 18).

[36] J. Mateos, D. Lavenier, M. Dimopoulou, A. Genot and M. Antonini. 'Primer design for DNA storage random access'. In: CORESA 2024 - 23ème conférence sur COmpression et REprésentation des Signaux Audiovisuels. Rennes, France, 2024, pp. 1–3. URL: https://hal.science/hal-04739300 (cit. on p. 19).

[37] N. Maurice, C. Frioux, C. Lemaitre and R. Vicedomini. 'Assessing assembly quality in metagenomes of increasing complexity sequenced with HiFi long reads'. In: 2024 - 24th Genome Informatics meeting. Hinxton, United Kingdom, Nov. 2024, pp. 1–21. URL: https://hal.science/hal-04822870 (cit. on p. 17).

[38] A. Regnier and P. Peterlongo. 'k-mer matrix compression'. In: SeqBIM 2024 - Journées annuelles du groupe de travail SeqBIM (Séquences en Bioinformatique, Informatique et Mathématiques). Rennes, France, 2024, pp. 1–1. URL: https://hal.science/hal-04855866.

[39] S. Romain, F. Legeai and C. Lemaitre. 'Understanding the limits of pangenome graphs for the analysis of large inversions in a complex of butterfly species'. In: 2024 - International Environmental and Agronomical Genomics symposium. Toulouse, France, 2024, pp. 1–1. URL: https://hal.science/hal-04721489 (cit. on p. 19).

[40] O. Sladký, P. Veselý and K. Břinda. 'Masked superstrings as a compact, indexable and dynamic representation of k-mer sets'. In: SeqBim 2024 - Journées sur les Séquences en Bioinformatique, Informatique et. Mathématiques. Rennes, France, Nov. 2024, pp. 1–3. URL: https://inria.hal.science/hal-04842867 (cit. on p. 12).

[41] T. K. M. Truong, R. Faure and R. Andonov. 'Assembling close strains in metagenome assemblies using discrete optimization'. In: BIOINFORMATICS 2024 - 15th International conference on bioinformatics models, methods and algorithms. Rome, Italy, 2024, pp. 1–10. URL: https://inria.hal.science/hal-04349675 (cit. on pp. 7, 16).

Doctoral dissertations and habilitation theses

[42] R. Faure. 'Haplotype assembly from long reads'. Université de Rennes; Université libre de Bruxelles, 27th Nov. 2024. URL: https://hal.science/tel-04818568 (cit. on pp. 16, 34).

[43] K. Hannoush. 'Dynamic Pan-genome Graphs'. Univ Rennes, Inria, CNRS, IRISA, France, 13th Dec. 2024. URL: https://inria.hal.science/tel-04861589 (cit. on pp. 13, 34).

[44] S. Romain. 'Identification, genotyping and representation of structural variants in pangenomes'. Université de Rennes, 8th Nov. 2024, pp. i270–i278. URL: https://theses.hal.science/tel-04825910 (cit. on p. 34).

[45] E. Roux. 'Food and Functionalities'. Université de Rennes, 29th Nov. 2024. URL: https://theses.hal.science/tel-04878765 (cit. on p. 34).

Reports & preprints

[46] S. Dubois, M. Zytnicki, C. Lemaitre and T. Faraut. *Pairwise graph edit distance characterizes the impact of the construction method on pangenome graphs*. 11th Dec. 2024. DOI: 10.1101/2024.12.06.627166. URL: https://inria.hal.science/hal-04871087 (cit. on p. 18).

[47] 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)*. 2024. URL: https://inria.hal.science/hal-04764171 (cit. on p. 12).

[48] O. Sladký, P. Veselý and K. Břinda. *Towards Efficient k-Mer Set Operations via Function-Assigned Masked Superstrings*. 11th Mar. 2024. DOI: 10.1101/2024.03.06.583483. URL: https://hal.science/hal-04573444 (cit. on p. 12).

[49] R. Vicedomini, F. Andreace, Y. Dufresne, R. Chikhi and C. Duitama González. *MUSET: Set of utilities for the construction of abundance unitig matrices from sequencing data*. 11th Dec. 2024. URL: https://hal.science/hal-04831168 (cit. on p. 17).

Other scientific publications

[50] C. Belliardo, N. Maurice, C. Frioux, C. Lemaitre, R. Vicedomini, S. Mondy, A. Péré, M. Bailly-Bechet and E. Danchin. 'Unlocking the Soil Microbiome: Unraveling Soil Microbial Complexity Using Long-Read Metagenomics'. In: EAGS 2024 - The International Environmental and Agronomical Genomics symposium. Toulouse, France, 2024, pp. 1–1. URL: https://hal.science/hal-04509213 (cit. on p. 21).

[51] C. Belliardo, S. Mondy, A. Pere, C. Lemaitre, R. Vicedomini, C. Frioux, D. J. Sherman, P. Abad, M. Bailly-Bechet and E. Danchin. 'Exploring and quantifying the soil genetic diversity captured by long and short-read shotgun metagenomic sequencing'. In: Journées 2024 du programme Agroécologie et Numérique. Rennes, France, 2024, pp. 1–1. URL: https://hal.science/hal-04423917 (cit. on p. 21).

[52] K. Břinda, Z. Iqbal and M. Baym. 'Phylogenetic compression as a core compression technique for extremely large microbial genome collections'. In: Genome Informatics 2024. Hinxton, Cambridge, United Kingdom, Nov. 2024. URL: https://inria.hal.science/hal-04842914 (cit. on p. 15).

[53] N. Maurice, C. Lemaitre, C. Frioux and R. Vicedomini. 'Metagenomic assembly of complex ecosystems with highly accurate long-reads'. In: Journées 2024 du PEPR Agroécologie et Numérique. Rennes, France, 2024, pp. 1–1. URL: https://inria.hal.science/hal-04425626 (cit. on p. 17).

Scientific popularization

[54] T. Lemane, M. Lescot and P. Peterlongo. 'Indexer les données génomiques : un défi de taille à relever'. In: *Interstices* (23rd Feb. 2024), pp. 1–5. URL: https://inria.hal.science/hal-04757321 (cit. on p. 36).

[55] P. Peterlongo and L. Robidou. 'Indexer des milliards d'éléments avec les filtres de Bloom'. In: *Interstices* (28th Feb. 2024), pp. 1–6. URL: https://inria.hal.science/hal-04570454 (cit. on p. 36).

12.3 Cited publications

[56] 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 (2016), pp. 1–14 (cit. on p. 15).

[57] P. Pandey, M. A. Bender, R. Johnson and R. Patro. 'A general-purpose counting filter: Making every bit count'. In: *Proceedings of the 2017 ACM international conference on Management of Data*. 2017, pp. 775–787 (cit. on pp. 13, 14).