UMR IRISA

Activity Report 2023

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

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, Researcher, HDR]
- Jacques Nicolas [INRIA, Senior Researcher, HDR]
- Riccardo Vicedomini [CNRS, Researcher, from Feb 2023]

Faculty Member

- Roumen Andonov [UNIV RENNES, Professor, HDR]

Post-Doctoral Fellow

- Loren Dejoies [INRIA, Post-Doctoral Fellow, from Feb 2023]

PhD Students

- Siegfried Dubois [INRAE, from Sep 2023]
- Victor Epain [INRIA, until Nov 2023]
- Roland Faure [UNIV RENNES]
- Garance Gourdel [UNIV RENNES, until Oct 2023]
- Khodor Hannoush [INRIA]
- Victor Levallois [INRIA, from Oct 2023]
- Nicolas Maurice [INRIA, from Oct 2023]
- Meven Mognol [UNIV RENNES , CIFRE]
- Lucas Robidou [INRIA, until Sep 2023]
- Sandra Romain [INRIA]

Technical Staff

- Gaëtan Benoit [INRIA, Engineer, from Mar 2023]
- Olivier Boulle [INRIA, Engineer]
- Anne Guichard [INRIA, Engineer, until Jan 2023]
- Julien Leblanc [CNRS, Engineer]
- Gildas Robine [CNRS, Engineer]
- Florestan de Moor [CNRS, Engineer, from Apr 2023]

Interns and Apprentices

- Siegfried Dubois [INRIA, Intern, until Jul 2023]
- Veronika Hendrychová [UNIV RENNES, Intern, from Sep 2023]
- Baptiste Hilaire [ENS PARIS-SACLAY, Intern, from Jun 2023 until Jul 2023]
- Arya Kaul [INRIA, Intern, from Sep 2023]
- Victor Levallois [INRIA, Intern, until Jun 2023]
- Nicolas Maurice [INRIA, Intern, until Jun 2023]
- Khac Minh Tam Truong [UNIV RENNES, Intern, from Mar 2023 until Jun 2023]

Administrative Assistant

- Marie Le Roic [INRIA]

Visiting Scientists

- Francesca Brunetti [UNIV SAPIENZA , from Nov 2023]
- Luca Parmigiani [UNIV POTSDAM, from May 2023 until Jul 2023]

External Collaborators

- Susete Alves Carvalho [INRAE, until Jan 2023]
- Erwan Drezen [INSTITUT PASTEUR, from Apr 2023]
- Fabrice Legeai [INRAE]
- Emeline Roux [UNIV RENNES]

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 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 main expertise is focused (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 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 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 matching 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 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.

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 ennemies 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 significative 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

On October 10, the "*L'Oréal Foundation and Unesco*" honored 35 young female doctoral and post-doctoral researchers with the Prix "*Jeunes Talents France 2023 Pour les Femmes et la Science*". Garance Gourdel, winner, was a doctoral student in the team. This prize highlights her research work that focuses on the use of algorithms to improve DNA reading, creating and analyzing new algorithms for processing and storing large volumes of data, such as that generated by sequencing. She defended her thesis in October at ENS Paris.

6.2 Results

We have presented an original method for Structural Variant genotyping in the prestigious international conference of bioinformatics ISMB/ECCB 2023. In addition to the talk, the paper was published in the journal Bioinformatics [20].

Advances in sequencing technologies have revealed the prevalence and importance of structural variations (deletions, duplications, inversions or rearrangements of DNA segments) which cover 5 to 10 times more bases in the genome than the point mutations commonly analyzed. Over the last 5 years, the rise of 3rd generation sequencing (long reads) has made it possible to characterise and catalogue the full range of SVs in many model organisms, such as in human. Then, the next step to fully understand variations in populations and associate them to phenotypes consists in assessing the presence or absence of known variants in numerous newly sequenced individuals. This is the genotyping problem. We proposed, in this work, SVjedi-graph, the first structural variant genotyping method dedicated to long read data and relying on a variation graph to represent all alleles of all variants in a single data structure. We showed that this graph model prevents the bias toward the reference alleles and allows maintaining high genotyping accuracy whatever the proximity of variants, contrary to other state of the art genotypers.

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 fimpera

Keywords: Indexation, Data structures, K-mer, Bloom filter, Bioinfirmatics search sequence, Search Engine

Functional Description: fimpera is a strategy for indexing and querying "hashtable-like" data structures named "AMQ" (for "Approximate Membership Query data structure"). When queried, those AMQs can yield false positives or overestimated calls. fimpera reduces their false positive rate by two orders of magnitude while reducing the overestimations. Without introducing any false negative, fimpera speeds up queries.

URL: https://github.com/lrobidou/fimpera

Publication: hal-03912993

Contact: Lucas Robidou

Participants: Lucas Robidou, Pierre Peterlongo

7.1.5 SVJedi-graph

Keywords: Structural Variation, Genotyping, High throughput sequencing, Sequence alignment

Functional Description: SVJedi-graph is a structural variation (SV) genotyper for long read data. It constructs a variation graph to represent all alleles of all SVs given as input. This graph-based approach allows to better represent close and overlapping SVs. Long reads are then aligned to this graph and the genotype of each variant is estimated based on allele-specific alignment counts. SVJedi-graph takes as input a variant file (VCF), a reference genome (fasta) and a long read file (fasta/fastq) and outputs the initial variant file with an additional column containing genotyping information (VCF).

URL: https://github.com/SandraLouise/SVJedi-graph

Contact: Claire Lemaitre

Participants: Claire Lemaitre, Sandra Romain

7.1.6 MTG-link

Keywords: Bioinformatics, Genome assembly, Barcode, Linked-reads, Gap-filling

Functional Description: MTG-Link is a local assembly tool dedicated to linked-read sequencing data. It leverages barcode information from linked-reads to assemble specific loci. Notably, the sequence to be assembled can be totally unknown (contrary to targeted assembly tools). It takes as input a set of linked-reads, the target flanking sequences and coordinates in GFA format and an alignment file in BAM format. It outputs the results in a GFA file.

Release Contributions: MTG-Link can now be used for various local assembly use cases, such as intra-scaffold and inter-scaffold gap-fillings, as well as the reconstruction of the alternative allele of large insertion variants. It is also directly compatible with the following linked-reads technologies, given that the barcodes are reported using the "BX:Z" tag: 10X Genomics, Haplotagging, stLFR and TELL-Seq.

URL: https://github.com/anne-gcd/MTG-Link

Publications: hal-03073966, hal-03074227, hal-03441914, hal-03886951

Contact: Claire Lemaitre

Participants: Anne Guichard, Fabrice Legeai, Claire Lemaitre

Partner: INRAE

7.1.7 PyRevSymG

Name: Python Reverse Symmetric Graph

Keywords: Directed graphs, Graph algorithmics, DNA sequencing

Functional Description: PyRevSymG is a Python3 API to store succession relationships between oriented fragments (in forward or reverse orientation) that have been sequenced from nucleotide sequence(s) in an oriented graph. For example, this API can be used for a genome assembly overlap-layout-consensus method.

URL: https://pypi.org/project/revsymg/

Contact: Victor Epain

Participant: Victor Epain

7.1.8 DnarXiv

Name: dnarXiv project platform

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

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

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

Contact: Olivier Boulle

Participants: Olivier Boulle, Dominique Lavenier

Partners: IMT Atlantique, Université de Rennes 1

7.1.9 MOF-SEARCH

Name: MOF-SEARCH

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

7.1.10 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

7.1.11 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

7.1.12 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.

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

Contact: Siegfried Dubois

Participants: Siegfried Dubois, Claire Lemaitre

Partner: INRAE

7.1.13 pancat

Name: PANgenome Comparison and Anlaysis 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.

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

Contact: Siegfried Dubois

Participants: Siegfried Dubois, Claire Lemaitre

Partner: INRAE

7.1.14 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

7.1.15 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

8 New results

8.1 Indexing data structures

8.1.1 Improvement of Approximate Membership Query data-structures with counts

Participants: Pierre Peterlongo, Lucas Robidou.

Approximate membership query (AMQ) data structures are widely used for indexing the presence of elements drawn from a large set. To represent the count of indexed elements, AMQ data structures can be generalized into "counting AMQ" data structures. This is for instance the case of the "counting Bloom filters". However, counting AMQ data structures suffer from false positive and overestimated calls. In this work we propose a novel strategy, called fimpera, that reduces the false-positive rate and overestimation rate of any counting AMQ data structure indexing k-mers (words of length k) from a set of sequences, along with their abundance. Applied on a counting Bloom filter, fimpera decreases its false-positive rate by an order of magnitude while reducing the number of overestimated calls. Furthermore, fimpera lowers the average difference between the overestimated calls and the ground truth. In addition, it slightly decreases the query time. fimpera does not require any modification of the original counting AMQ data structure, it does not generate false-negative calls, and causes no memory overhead. The unique drawback is that fimpera yields a negligible amount of a new kind of false positives and overestimated calls [18].

8.1.2 Indexing large sets of sequencing data

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 developped kmindex (7.1.2), 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. We demonstrate 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 [37].

8.1.3 Phylogenetic compression

Participants: Karel Břinda.

Comprehensive collections approaching millions of sequenced genomes have become central information sources in the life sciences. However, the rapid growth of these collections makes it effectively impossible to search these data using tools such as BLAST and its successors. To address this challenge, we developed a technique called phylogenetic compression, which uses evolutionary history to guide compression and efficiently search large collections of microbial genomes using existing algorithms and data structures [32]. We showed that, when applied to modern diverse collections approaching millions of genomes, lossless phylogenetic compression improves the compression ratios of assemblies, de Bruijn graphs, and *k*-mer indexes by one to two orders of magnitude (implemented in a software tool called MiniPhy (7.1.10)). Additionally, we developed a pipeline called MOF-Search (7.1.9) for a BLAST-like search over these phylogeny-compressed reference data, and demonstrated it can align genes, plasmids, or entire sequencing experiments against all sequenced bacteria until 2019 on ordinary desktop computers within a few hours. Phylogenetic compression has broad applications in computational biology and may provide a fundamental design principle for future genomics infrastructure [32].

8.1.4 Processing compressed genomic data

Participants: Roland Faure, Baptiste Hilaire, Dominique Lavenier.

Efficiently managing large DNA datasets necessitates the development of highly effective sequence compression techniques to reduce storage and computational requirements. Here, we explore the potential of a lossy compression technique, Mapping-friendly Sequence Reductions (MSRs) which is a generalization of homopolymer compression to improve the accuracy of alignment tools. Essentially, MSRs deterministically transform sequences into shorter counterparts, in such a way that if an original query and a target sequence align, their reduced forms will align as well. While homopolymer compression is one example of an MSR, numerous others exist. These rapid computations yield lossy representations of the original sequences. Notably, the reduced sequences can be stored, aligned, assembled, and indexed much like regular sequences. MSRs could be used to improve the efficiency of taxonomic classification tools, by indexing and querying reduced sequences. Our experimentation with a mixture of 10 *E. coli* strains, demonstrates that this approach can yield greater precision than indexing and querying a reduced portion of k-mers. Other tasks could benefit from sequence reduction, such as mapping, genome assembly, and structural variant detection [26].

8.1.5 K-mer-based methods for ancient oral metagenomics

Participants: Riccardo Vicedomini.

Despite the implementation of strict laboratory protocols for controlling ancient DNA contamination, samples still remain highly vulnerable to environmental contamination. Such contamination can significantly alter microbial composition, leading to inaccurate conclusions in downstream analyses. Within the co-supervision of a PhD student at the Institut Pasteur (Paris, France), we contributed to the work related to two *k*-mer-based methods for addressing the following two challenges in ancient metagenomics: (i) microbial source tracking for contamination assessment [13] and (ii) read-level contamination removal [12].

The first method is based on the construction of a k-mer matrix [8] which stores the presence/absence of k-mers across multiple samples of different well-characterised sources. Such a matrix is then used to predict the proportion of each source in unknown input samples.

The second method allows to retain, from an input contaminated set, reads likely to belong to a specific source of interest. On synthetic data, it achieves over 89.53% sensitivity and 94.00% specificity. On real datasets, aKmerBroom shows higher read retainment (+60% on average) than competing methods.

8.2 Theoretical studies of string algorithms

8.2.1 Optimal Square Detection Over General Alphabets

Participants: Garance Gourdel.

Squares (fragments of the form xx, for some string x) are arguably the most natural type of repetition in strings. The basic algorithmic question concerning squares is to check if a given string of length n is square-free, that is, does not contain a fragment of such form. We show that testing square-freeness of a length-n string over general alphabet of size σ can be done with $O(n \log \sigma)$ comparisons, and cannot be done with $o(n \log \sigma)$ comparisons. This result comes with an $O(n \log \sigma)$ time algorithm in the Word RAM model [21].

8.2.2 Compressed Indexing for Consecutive Occurrences

Participants: Garance Gourdel.

The fundamental question considered in algorithms on strings is that of indexing, that is, preprocessing a given string for specific queries. By now we have a number of efficient solutions for this problem when the queries ask for an exact occurrence of a given pattern P. However, practical applications motivate the necessity of considering more complex queries, for example concerning near occurrences of two patterns.

Recently, Bille et al. [45] introduced a variant of such queries, called gapped consecutive occurrences, in which a query consists of two patterns P_1 and P_2 and a range $[a, b]$, and one must find all consecutive occurrences (q_1, q_2) of P_1 and P_2 such that $q_2 - q_1 \in [a, b]$. By their results, we cannot hope for a very efficient indexing structure for such queries, even if $a = 0$ is fixed (although at the same time they provided a non-trivial upper bound). Motivated by this, we focus on a text given as a straight-line program (SLP) and design an index taking space polynomial in the size of the grammar that answers such queries in time optimal up to polylog factors [22, 35].

8.2.3 Masked superstrings

Participants: Karel Břinda.

The popularity of k-mer-based methods has recently led to the development of compact k-mer-set representations, such as simplitigs/Spectrum-Preserving String Sets (SPSS), matchtigs, and eulertigs. These aim to represent k-mer sets via strings that contain individual k-mers as substrings more efficiently than the traditional unitigs. We demonstrated that all such representations can be viewed as superstrings of input k-mers, and as such can be generalized into a unified framework that we call the masked superstring of k-mers [39]. We studied the complexity of masked superstring computation and proved NP-hardness for both k-mer superstrings and their masks [39]. We then designed local and global greedy heuristics for efficient computation of masked superstrings, implemented them in a program called KmerCamel (7.1.11), and evaluate their performance using selected genomes and pan-genomes [39]. Overall, masked superstrings unify the theory and practice of textual k-mer set representations and provide a useful framework for optimizing representations for specific bioinformatics applications.

8.3 Algorithms for genome assembly

8.3.1 Scaffolding step in genome assembly

Participants: Victor Epain, Rumen Andonov, 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 distances 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 distances 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 [34].

8.3.2 Local assembly with linked-read data

Participants: Anne Guichard, Fabrice Legeai, Claire Lemaitre.

Local assembly consists in reconstructing a sequence of interest from a sample of sequencing reads without having to assemble the entire genome, which is time and labor intensive. This is particularly useful when studying a *locus* of interest, for gap-filling in draft assemblies, as well as for alternative allele reconstruction of large insertion variants. Whereas linked-read technologies have a great potential to assemble specific *loci* as they provide long-range information, while maintaining the power and accuracy of short-read sequencing, there is a lack of local assembly tools for linked-read data.

We present MTG-Link (7.1.6), a novel local assembly tool dedicated to linked-reads. The originality of the method lies in its read subsampling step which takes advantage of the barcode information contained in linked-reads mapped in flanking regions of each targeted locus. Our approach relies then on our tool MindTheGap [10] to perform local assembly of each locus with the read subsets. MTG-Link tests different parameters values for gap-filling, followed by an automatic qualitative evaluation of the assembly.

We validated our approach on several datasets from different linked-read technologies. We show that MTG-Link is able to successfully assemble large sequences, up to dozens of Kb. We also demonstrate that the read subsampling step of MTG-Link considerably improves the local assembly of specific *loci* compared to other existing short-read local assembly tools. Furthermore, MTG-Link was able to fully characterize large insertion variants in a human genome and improved the contiguity of a 1.3 Mb *locus* of biological interest in several individual genomes of the mimetic butterfly *Heliconius numata* [15].

8.3.3 Separating strains in metagenome assemblies with long reads

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

Long read assemblers struggle to distinguish closely related strains of the same species and collapse them into a single sequence. This is very limiting when analysing a metagenome, as different strains can have important functional differences. We have designed a new methodology supported by a software called HairSplitter (7.1.15), which recovers the strains from a strain-oblivious assembly and long reads. The originality of the method lies in a custom variant calling step that works with erroneous reads

and separates an unknown number of haplotypes. On simulated datasets, we show that HairSplitter significantly outperforms the state of the art when dealing with metagenomes containing many strains of the same species [25, 40]

We also propose an alternative approach for the strain separation problem using Integer Linear Programming (ILP). 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 sequencing 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. We think the new methodological ideas presented in this paper will help democratizing strain-separated assembly [27].

8.4 Algorithms for variant detection and analysis

8.4.1 Structural Variation genotyping with variant graphs

Participants: Claire Lemaitre, Sandra Romain.

One of the problems in Structural Variant (SV) analysis is the genotyping of variants. It consists in estimating the presence or absence of a set of known variants in a newly sequenced individual. Our team previously released SVJedi, one of the first SV genotypers dedicated to long read data. The method is based on linear representations of the allelic sequences of each SV. While this is very efficient for distant SVs, the method fails to genotype some closely located or overlapping SVs. To overcome this limitation, we present a novel approach, SVJedi-graph (7.1.5), which uses a sequence graph instead of linear sequences to represent the SVs.

In our method, we build a variation graph to represent in a single data structure all alleles of a set of SVs. The long reads are mapped on the variation graph and the resulting alignments that cover allele-specific edges in the graph are used to estimate the most likely genotype for each SV. Running SVJedi-graph on simulated sets of close and overlapping deletions showed that this graph model prevents the bias toward the reference alleles and allows maintaining high genotyping accuracy whatever the SV proximity, contrary to other state of the art genotypers. On the human gold standard HG002 dataset, SVJedi-graph obtained the best performances, genotyping 99.5% of the high confidence SV callset with an accuracy of 95% in less than 30 min [20].

8.4.2 Towards an edit distance between pangenome graphs

Participants: Siegfried Dubois, Claire Lemaitre.

A variation graph is a data structure that aims to represent variations among a collection of genomes. It is a sequence graph where each genome is embedded as a path in the graph with the successive nodes, along the path, corresponding to successive segments on the associated genome sequence. Shared subpaths correspond to shared genomic regions between the genomes and divergent path to variations: this structure features inversions, insertions, deletions and substitutions. 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 such as those implemented in the state-of-the-art pangenome graph builders minigraph-cactus and pggb. The question that arises is to what extent the construction method influences the resulting graph and therefore to what extent the resulting graph reflects genuine genomic variations. We propose to address this question by constructing an edition script between two variation graphs built from the same set of genomes which provides a measure of similarity, and more importantly that enables to identify discordant regions between the two graphs. We proceed by comparing, for each genome, the two corresponding paths in the two graphs which correspond to two possibly different segmentations of the same genomic sequence. As such, for each

interval defined by the nodes of the path of the genome in the first graph, we define a set of relations with the nodes of the second graph, such as equalities, prefix and suffix overlaps... which allows for a calculation of how many elementary operations, such as fusions and divisions of nodes, are required to go from one graph to another. We tested our method on variation graphs constructed using both simulated dataset as well as a real dataset made of 15 yeast telomere-to-telomere phased genome assemblies, with minigraph-cactus and pggb as the graph construction tools. We showed that two graphs built with the same tool, minigraph-cactus, but with different incorporation orders of genomes can be more different from one another than two graphs built with the two different tools. We also showed thar our distance allows to pinpoint and vizualize the specific areas of the graph and genomes that are impacted by the changes in segmentation. The method is implemented in a Python tool named Pancat (7.1.13) [33, 24].

8.4.3 Efficient detection of positive selection in large population genomic datasets

Participants: Pierre Peterlongo.

Genomic regions under positive selection harbor variation linked for example to adaptation. Most tools for detecting positively selected variants have computational resource requirements rendering them impractical on population genomic datasets with hundreds of thousands of individuals or more. We have developed and implemented an efficient haplotype-based approach able to scan large datasets and accurately detect positive selection. We achieve this by combining a pattern matching approach based on the positional Burrows–Wheeler transform with model-based inference which only requires the evaluation of closed-form expressions. We evaluate our approach with simulations, and find it to be both sensitive and specific. The computational resource requirements quantified using UK Biobank data indicate that our implementation is scalable to population genomic datasets with millions of individuals. Our approach may serve as an algorithmic blueprint for the era of “big data” genomics: a combinatorial core coupled with statistical inference in closed form [16].

8.5 Information storage on DNA molecules

8.5.1 Encoding data under biological and indexing constraints

Participants: Dominique Lavenier.

We have developed a method based on a dynamic sliding window encoding (DSWE) for storing encrypted data in a DNA form taking into account biological constraints and prohibited nucleotide motifs used for data indexing. Its originality is twofold. First, it takes advantage of variable length DNA codewords to avoid homopolymers longer than N bases when encoding binary data. Second, it relies on a sliding window to prevent the creation of prohibited motifs of nucleotides, adding non-coding bases when necessary. Contrarily to existing schemes, scaling DSWE to high values of N and of numbers of prohibited motifs is extremely simple. It is furthermore independent of the cryptosystem. We provide the theoretical information rate of our proposal for a given number of prohibited motifs and a maximum homopolymer length. Experiments show that in general, it offers much higher performances than existing schemes. [23].

8.5.2 In vitro construction of long artificial DNA molecule for encoding numerical data

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

In absence of DNA template, the ab initio production of long double-stranded DNA molecules of predefined sequences is particularly challenging. The DNA synthesis step remains a bottleneck for

many applications such as functional assessment of ancestral genes, analysis of alternative splicing or DNA-based data storage. We worked on a fully in vitro protocol to generate very long double-stranded DNA molecule starting from commercially available short DNA blocks in less than 3 days. This innovative application of Golden Gate assembly allowed us to streamline the assembly process to produce a 24 kb long DNA molecule storing part of the Universal Declaration of Human rights and citizens. The DNA molecule produced can be readily cloned into suitable host/vector system for amplification and selection. [36].

8.6 Processing-in-Memory

Participants: Charles Deltel, Florestan de Moor, Erwan Drezen, Dominique Lavenier, Meven Mognol, Gildas Robine.

Processing-in-Memory (PIM) consists of processing capabilities tightly coupled with the main memory. Contrary to bringing all data into a centralized processor, which is far away from the data storage, in-memory computing processes the data directly where it resides, suppressing most data movements, and, thereby greatly improving the performance of massive data applications by orders of magnitude. NGS data analysis completely falls in these application domains where PIM can strongly accelerate the main time-consuming software in genomic and metagenomic areas. More specifically, mapping algorithms, intensive sequence comparison algorithms or bank searching, for example, can highly benefit of the parallel nature of the PIM concept.

8.6.1 Data structures

In the framework of the European BioPIM project, we have studied (from a parallelism point of view) a number of data structures used extensively in genomics software to assess the benefits of implementing them on PIM architectures. The following data structures have been studied: bloom filters, Burrows–Wheeler transform and hash tables. A detailed report on the evaluations is currently available on the GenoPIM project website.

8.6.2 Programing Environment

The programming model for processing-in-memory is not yet well defined. The question of automatic (or semi-automatic) parallelization remains open. The tools available for programming a complete application on a PIM-equipped architecture are relatively low-level. We are working on the design of a C++ programming environment to unify the programming of CPU and PIM memory processing.

8.6.3 Applications

Memory components based on PIM principles have been developed by UPMEM, a young startup founded in 2015. The company has designed an innovative DRAM processing unit (DPU), a RISC processor integrated directly into the memory chip, on the DRAM die. We are using a UPMEM PIM server equipped with 160 GB of PIM memory and 256 GB of legacy memory to carry out full-scale experiments on the implementation of several genomic software for long DNA comparison, bacterial genome comparison, protein sequence alignment, data compression and sorting algorithms.

Initial experiments show that, for certain applications, it is possible to achieve a 20-fold acceleration compared with standard multicore platforms.

8.7 Benchmarks and Reviews

8.7.1 Benchmarking metagenome assemblers for long reads and application to soil microbiomes

Participants: Claire Lemaitre, Nicolas Maurice, Riccardo Vicedomini.

We developed an automated pipeline, Mapler (7.1.14), to assess the performances of long read metagenome assemblers, with a special focus on tools dedicated to high fidelity long reads such as PacBio HiFi reads. It compares five assembly tools, namely metaMDBG, metaflye, Hifiasm-meta, opera-ms and miniasm, and uses various evaluation metrics such as reference-based, reference-free and binning-based evalutation metrics. We applied this pipeline on several real metagenome datasets of various complexities, including publicly available mock communities with well characterized species contents and tunnel culture soil metagenomes with high and unknown species complexities. We showed that the different evaluation metrics are complementary and that high fidelity long read data allows drastic improvements in the number of obtained good quality Metagenome Assembled Genomes (MAGs) with respect to low-fidelity long reads or short read data, with MetaMDBG out-performing other HiFi-dedicated assemblers. However, for soil metagenomes most reconstructed MAGs remain of low quality because of the very large number of species in these microbiomes and the probable under-sampling of this diversity by this type of sequencing [41].

8.7.2 Introduction to bioinformatics methods for metagenomic and metatranscriptomic analyses

Participants: Claire Lemaitre.

In this book chapter, we review the different bioinformatics analyses that can be performed on metagenomics and metatranscriptomics data. We present the differences of this type of data compared to standard genomics data and highlight the methodological challenges that arise from it. We then present an overview of the different methodological approaches and tools to perform various analyses such as taxonomic annotation, genome assembly and binning and *de novo* comparative genomics [28].

8.8 Bioinformatics Analysis

8.8.1 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. We participated in the genome assembly and analyses of some major agricultural pests or their natural ennemies. In particular, we performed a genome-wide identification of lncRNAs associated with viral infection in the lepidopteran pest *Spodoptera frugiperda* [19] and participated in a detailed study of a genomic region linked to reproductive mode variation in the pea aphid [17].

In most cases, the genomes and their annotations were hosted in the BIPAA information system, allowing collaborative curation of various sets of genes and leading to novel biological findings [42].

8.8.2 First chromosome scale genomes of ithomiine butterflies

Participants: Fabrice Legeai, Claire Lemaitre.

In the framework of a former ANR project (SpecRep 2014-2019), we worked on *de novo* genome assembly of several ithomiine butterflies. Due to their high heterozygosity level and to sequencing data of various quality, this was a challenging task and we tested numerous assembly tools. Finally, this work led to the generation of high-quality, chromosome-scale genome assemblies for two *Melinaea* species, *M. marsaeus* and *M. menophilus*, and a draft genome of the species *Ithomia salapia*. We obtained genomes with a size ranging from 396 Mb to 503 Mb across the three species and scaffold N50 of 40.5 Mb and 23.2

Mb for the two chromosome-scale assemblies. Various genomics and comparative genomics analyses were performed and revealed notably independent gene expansions in ithomiines and particularly in gustatory receptor genes.

These three genomes constitute the first reference genomes for the ithomiine butterflies (Nymphalidae: Danainae), which represent the largest known radiation of Müllerian mimetic butterflies and dominate by number the mimetic butterfly communities. This is therefore a valuable addition and a welcome comparison to existing biological models such as Heliconius, and will enable further understanding of the mechanisms of mimetism and adaptation in butterflies [14].

8.8.3 The Silene latifolia genome and its giant Y chromosome

Participants: Claire Lemaitre.

In some species, the Y is a tiny chromosome but the dioecious plant *Silene latifolia* has a giant 550 Mb Y chromosome, which has remained unsequenced so far. We participated in a collaborative project that sequenced and obtained a high-quality male *S. latifolia* genome. We participated in particular in the comparative analysis of the sex chromosomes with outgroups, that showed that the Y is surprisingly rearranged and degenerated for a 11 MY-old system. Recombination suppression between X and Y extended in a stepwise process, and triggered a massive accumulation of repeats on the Y, as well as in the non-recombining pericentromeric region of the X, leading to giant sex chromosomes [38].

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

Francesca Brunetti

Status PhD student

Institution of origin: Sapienza University of Rome

Country: Italy

Dates: Starting from November 2023

Context of the visit: Development of applications of phylogenetic compression for biology and public health

Mobility program/type of mobility: Research stay (1 year)

Luca Parmigiani

Status PhD student

Institution of origin: Universität Bielefeld

Country: Germany

Dates: 1st of May 2023 to 31

Context of the visit: Alpaca Secondment

Mobility program/type of mobility: Research stay

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, 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), Netherlands

Inria contact: Pierre Peterlongo

Coordinator:

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 BioPiM project

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

Duration: From May 1, 2022 to April 30, 2026

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.

10.2.2 Other european programs/initiatives

Partnership Hubert Curien (PHC) PROCOPE 2023

Participants: Rumen Andonov, Karel Břinda, Roland Faure, Sandra Romain, Riccardo Vicedomini.

Title: Novel Reference Structures for Long-Read Dataset Search (RESLO)

Duration: From Jan 2023 to Dec 2024

Partners: Heinrich Heine University Düsseldorf

Description: The RESLO project aims to develop a new reference structure for representing PacBio sequencing experiments in search engines, addressing the challenges of managing large and diverse genomic data. By generalizing de Bruijn graphs, the team seeks to create a graphical representation that effectively indexes various PacBio sequencing technologies and protocols. The focus is on developing methods to capture critical signals like haplotype structure, repeat structure, and structural variants. This advancement will enhance the comparability and searchability of genomic data, significantly benefiting biology, epidemiology, and medicine.

Coordinators: Karel Břinda, Gunnar Klau

Inria Contact: Karel Břinda

10.3 National initiatives

10.3.1 PEPR

Project MolecularXiv. Targeted Project 2: From digital data to bases

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.

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 developping 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 and taking advantages from 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 Supergene: The consequences of supergene evolution

Participants: Anne Guichard, Dominique Lavenier, Fabrice Legeai, Claire Lemaitre, Pierre Peterlongo.

- Coordinator: M. Joron (Centre d'Ecologie Fonctionnelle et Evolutive (CEFE) UMR CNRS 5175, Montpellier)
- Duration: 54 months (Nov. 2018 – Apr. 2023)
- Partners: CEFE (Montpellier), MNHN (Paris), Genscale Inria/IRISA Rennes.
- Description: The Supergene project aims at better understanding the contributions of chromosomal rearrangements to adaptive evolution. Using the supergene locus controlling adaptive mimicry in a polymorphic butterfly from the Amazon basin (*H. numata*), the project will investigate the evolution of inversions involved in adaptive polymorphism and their consequences on population biology. GenScale's task is to develop new efficient methods for the detection and genotyping of inversion polymorphism with several types of re-sequencing data.

Project SeqDigger: Search engine for genomic sequencing data

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

- Coordinator: P. Peterlongo
- Duration: 48 months (jan. 2020 – Dec. 2024)
- 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: https://www.cesgo.org/seqdigger/

Project Divalps: diversification and adaptation of alpine butterflies along environmental gradients

Participants: Fabrice Legeai, Claire Lemaitre, Sandra Romain.

- Coordinator: L. Desprès (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: Charles Deltel, Florestan De Moor, Dominique Lavenier, Meven Mognol, Gildas Robine.

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: https://genopim.irisa.fr/

10.3.3 Inria Exploratory Action

Défi Inria OmicFinder

Participants: Pierre Peterlongo, Victor Levallois.

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: https://project.inria.fr/omicfinder/

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

Participants: Gaëtan Benoit, 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

10.4.1 LABEX CominLabs

dnarXiv: Archiving Information on DNA Molecules

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

- Coordinator: Dominique Lavenier
- Duration: 48 months (2021-2024)
- Web site: dnarXiv Project
- 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.

11 Dissemination

11.1 Promoting scientific activities

Participants: Karel Břinda, Victor Epain, Garance Gourdel, Dominique Lavenier, Fabrice Legeai, Claire Lemaitre, Jaques Nicolas, Pierre Peterlongo.

11.1.1 Scientific events: selection

Chair of conference program committees

- seqBIM2023: national meeting of the sequence algorithms GT seqBIM, Lille, Nov 2023 (70 participants, 2 days) [C. Lemaitre]

Member of the conference program committees

- ISMB/ECCB 2023: 31th Annual Conference on Intelligent Systems for Molecular Biology and 22nd Annual European Conference on Computational Biology, Lyon, 2023 [C. Lemaitre, D. Lavenier]
- DIMES 2023: Workshop on Disruptive Memory Systems, Koblenz, Germany, 2023 [D. Lavenier]

11.1.2 Journal

Member of the editorial boards

- Insects [F. Legeai]

Reviewer - reviewing activities

- Genome Biology [C. Lemaitre]
- Genome Research [P. Peterlongo, K. Břinda]
- Bioinformatics [D. Lavenier]
- BMC Bioinformatics [D. Lavenier]
- BMC Genomics [F. Legeai]
- Computational and Structural Biotechnology Journal [D. Lavenier]
- PCI Math Comp Biol [P. Peterlongo]
- Methods [D. Lavenier]
- Scientific Reports [D. Lavenier]
- Nature Scientific Data [F. Legeai]
- Wellcome Open Research [F. Legeai]

11.1.3 Invited talks

- D. Lavenier, "Exploring genomic algorithms on Processing-in-Memory Architecture", 5th workshop on Accelerator Architecture for Computational Biology and Bioinformatics, ISCA 2023, Orlando, Florida, June 2023
- P. Peterlongo, Keynote, "Indexing Large Metagenomic Projects. Application to the Tara Oceans Datasets", WCTA 2023, Pisa Italy, September 2023.
- K. Břinda, "Rapid inference of antibiotic resistance and susceptibility by Genomic Neighbor Typing", Institute of Medical Microbiology, University Hospital of Düsseldorf, Germany, June 2023.
- K. Břinda, "Under the hood: The role of string algorithms in contemporary biological research", CPM 2023 Summer School, ENS Paris, France, June 2023.
- K. Břinda, "Efficient and robust search of microbial genomes via phylogenetic compression", joint Alpaca/Pangaia consortia seminar, online, October 2023.
- K. Břinda, "K-mers and their graphs in computational biology", Graphs and Bioinformatics 2023, Sorbonne University, Paris, France, November 2023.
- K. Břinda, "Efficient search of microbial genomes via phylogenetic compression", Bioinformatics seminar, Faculty of Mathematics and Physics, Charles University in Prague, Czech Republic, December 2023.
- K. Břinda, "Towards diagnostics of antibiotic resistance within an hour", DIANA Biotechnologies, Prague, Czech Republic, December 2023.

11.1.4 Leadership within the scientific community

- Members of the Scientific Advisory Board of the GDR BIM (National Research Group in Molecular Bioinformatics) [P. Peterlongo, C. Lemaitre]
- Animator of the Sequence Algorithms axis (seqBIM GT) of the BIM and IM GDRs (National Research Groups in Molecular Bioinformatics and Informatics and Mathematics respectively) (170 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]

11.1.5 Scientific expertise

- Scientific expert for the DGRI (Direction générale pour la recherche et l'innovation) from the Ministère de l'Enseignement Supérieur, de la Recherche et de l'Innovation (MESRI) [D. Lavenier]

11.1.6 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]
- Institutional delegate representative of INRIA in the GIS BioGenOuest regrouping all public research platforms in Life Science in the west of France (régions Bretagne/ Pays de Loire) [J. Nicolas]
- 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]
- Recruitment committees: member of the CR/ISFP recruitment committee of the Rennes Inria center [C. Lemaitre], president of the recruitment jury of an INRAe engineer [F. Legeai]
- Members of the board of the Brittany doctoral students’ association Nicomaque [V. Epain, G. Gourdel]

11.2 Teaching - Supervision - Juries

Participants: Rumen Andonov, Karel Břinda, Siegfried Dubois, Roland Faure, Dominique Lavenier, Fabrice Legeai, Claire Lemaitre, Jaques Nicolas, Pierre Peterlongo, Emeline Roux.

11.2.1 Teaching administration

- In charge of the master’s degree "Nutrition Sciences des Aliments" (NSA) of University of Rennes (72 students) [E. Roux]

11.2.2 Teaching

- Licence : R. Andonov, Models and Algorithms in Graphs, 100h, L3, Univ. Rennes, France.
- Licence : E. Roux, Biochemistry, 90h, L1 and L3, Univ. Rennes, France.
- Licence : S. Dubois, Java, 21h, L1 ISTN, Univ. Rennes, France.
- Master : R. Andonov, Modeling and Discrete Optimisation (MOD), 82h, M1 Miage, Univ. Rennes, France.
- Master : R. Faure, Object Oriented Programming (OOP), 32h, M1, Univ. Rennes, France
- Master : C. Lemaitre, P. Peterlongo, Algorithms on Sequences, 52h, M2, Univ. Rennes, France.
- Master : R. Vicedomini, K. Břinda, Experimental Bioinformactics, 24h, M1, ENS Rennes, France.
- Master : F. Legeai, RNA-Seq, Metagenomics and Variant discovery, 10h, M2, National Superior School Of Agronomy, Rennes, France.
- Master : E. Roux, biochemistry, 120h, M1 and M2, Univ. Rennes, France.

11.2.3 PhD Supervision

- PhD: G. Gourdel, Sketch-based approaches to process massive string data [30], defended: 26/10/2023, P. Peterlongo.
- PhD: V. Epain, Assemblage de fragments ADN : structures de graphes et scaffolding de génomes de chloroplastes, [29] defended: 27/11/2023, R. Andonov, D. Lavenier, JF Gibrat.
- PhD: L. Robidou, Search engine for genomic sequencing data , defended 21/09/2023, P. Peterlongo. [31]
- PhD in progress: S. Romain, Genome graph data structures for Structural Variation analyses in butterfly genomes, since 01/09/2021, C. Lemaitre, F. Legeai.
- PhD in progress: K. Hannoush, Pan-genome graph update strategies, since 01/09/2021, P. Peterlongo, C. Marchet.

- PhD in Progress: R. Faure, Recovering end-to-end phased genomes, since 01/10/2021, D. Lavenier, J-F. Flot.
- PhD in progress: M. Mognol, Processing-in-Memory, since 01/04/2022, D. Lavenier.
- PhD in progress: S. Dubois, Characterizing structural variation in pangenome graphs, 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, since 01/10/2023, C. Lemaitre, C. Frioux, R. Vicedomini.
- PhD in progress: V. Levallois, Indexing genomic data, since 01/10/2023, P. Peterlongo.
- PhD in progress: C. Duitama (Sequence Bioinformatics group, Institut Pasteur, Paris), Algorithms based on k-mers for ancient oral metagenomics, since 07/10/2020, R. Chikhi, H. Richard, R. Vicedomini.

11.2.4 Juries

- *Member of HDR thesis jury*
  - Mikael Salson, Univ. Lille, Nov. 2023 [D. Lavenier]
- *President of PhD thesis jury*
  - L. Robidou, Univ. Rennes, Sept 2023 [J. Nicolas]
- *Member of PhD thesis jury*
  - Eva Gil San Antonio, Univ. Nice, March 2023 [D. Lavenier]
  - Sebastian Schmidt, Univ. Helsinki, Finland, August 2023 [P. Peterlongo]
- *Member of PhD thesis committee*
  - Francesco Andreace, Pasteur Paris, [P. Peterlongo]
  - Thomas Baudeau, Univ. Lille [C. Lemaitre]
  - Antoine Dequay, Univ. Rennes [D. Lavenier]
  - Aref Ezzeddine, IMT-A, [D. Lavenier]
  - Khodor Hannoush, Univ. Rennes [K. Břinda]
  - Léo de La Fuente, Univ. Rennes [D. Lavenier]
  - Maël Lefeuvre, MNHN [J. Nicolas]
  - Xavier Pic, Univ. Nice [D. Lavenier]
  - Léa Vandamme, Univ. Lille, [P. Peterlongo]

11.3 Popularization

Participants: Victor Epain, Dominique Lavenier, Claire Lemaitre, Pierre Peterlongo.

11.3.1 Internal or external Inria responsibilities

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

11.3.2 Articles and contents

- Popularization article in Interstices "Le SARS-CoV-2 : une expérience inédite de surveillance génomique mondiale" (link) [44] [C. Lemaitre]
- Popularization article in Interstices "Stocker les données : la piste prometteuse de l'ADN" (link) [43] [D. Lavenier]

11.3.3 Education

- Conferences IUT Genie Biologique IUT Quimper, "Stockage de données sur ADN" [D. Lavenier]

11.3.4 Interventions

- Chiche! Interventions in high school classes to make high school students aware of research careers in the digital sector. Three interventions made in 2023 [P. Peterlongo]
- Organisation and animation of the Inria 2023 Ethics Workshop by Nicomaque [V. Epain]

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.

[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.

[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.

[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.

[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.

[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.

[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.

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

[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.

[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.

[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.

12.2 Publications of the year

International journals

[12] C. Duitama González, S. Rangavittal, R. Vicedomini, R. Chikhi and H. Richard. 'aKmerBroom: Ancient oral DNA decontamination using Bloom filters on k-mer sets'. In: *iScience* 26.11 (Nov. 2023), p. 108057. DOI: 10.1016/j.isci.2023.108057. URL: https://hal.sorbonne-universite.fr/hal-04303607.

[13] C. Duitama González, R. Vicedomini, T. Lemane, N. Rascovan, H. Richard and R. Chikhi. 'decOM: similarity-based microbial source tracking of ancient oral samples using k-mer-based methods'. In: *Microbiome* 11.1 (6th Nov. 2023), pp. 1–12. DOI: 10.1186/s40168-023-01670-3. URL: https://hal.science/hal-04303600.

[14] J. Gauthier, J. Meier, F. Legeai, M. McClure, A. Whibley, A. Bretaudeau, H. Boulain, H. Parrinello, S. T. Mugford, R. Durbin, C. Zhou, S. McCarthy, C. W. Wheat, F. Piron-Prunier, C. Monsempes, M.-C. François, P. Jay, C. Noûs, E. Persyn, E. Jacquin-Joly, C. Meslin, N. Montagné, M. Elias and C. Lemaitre. 'First chromosome scale genomes of ithomiine butterflies (Nymphalidae: Ithomiini): comparative models for mimicry genetic studies'. In: *Molecular Ecology Resources* (2023), pp. 1–14. DOI: 10.1111/1755-0998.13749. URL: https://inria.hal.science/hal-03926527.

[15] A. Guichard, F. Legeai, D. Tagu and C. Lemaitre. 'MTG-Link: leveraging barcode information from linked-reads to assemble specific loci'. In: *BMC Bioinformatics* 24.1 (14th July 2023), p. 284. DOI: 10.1186/s12859-023-05395-w. URL: https://inria.hal.science/hal-04166273.

[16] B. Kirsch-Gerweck, L. Bohnenkämper, M. T. Henrichs, J. N. Alanko, H. Bannai, B. Cazaux, P. Peterlongo, J. Burger, J. Stoye and Y. Diekmann. 'HaploBlocks: Efficient Detection of Positive Selection in Large Population Genomic Datasets'. In: *Molecular Biology and Evolution* 40.3 (4th Mar. 2023), pp. 1–12. DOI: 10.1093/molbev/msad027. URL: https://inria.hal.science/hal-04351491.

[17] M. Rimbault, F. Legeai, J. Peccoud, L. Mieuzet, E. Call, P. Nouhaud, H. Defendini, F. Mahéo, W. Marande, N. Théron, D. Tagu, G. Le Trionnaire, J.-C. Simon and J. Jaquiéry. 'Contrasting Evolutionary Patterns Between Sexual and Asexual Lineages in a Genomic Region Linked to Reproductive Mode Variation in the pea aphid'. In: *Genome Biology and Evolution* 15.9 (Sept. 2023), evad168. DOI: 10.1093/gbe/evad168. URL: https://inria.hal.science/hal-04350785.

[18] L. Robidou and P. Peterlongo. 'fimpera: drastic improvement of Approximate Membership Query data-structures with counts'. In: *Bioinformatics* 39.5 (1st May 2023), pp. 1–16. DOI: 10.1101/2022.06.27.497694. URL: https://inria.hal.science/hal-03912993.

[19] S. Robin, F. Legeai, V. Jouan, M. Ogliastro and I. Darboux. 'Genome-wide identification of lncRNAs associated with viral infection in Spodoptera frugiperda'. In: *Journal of General Virology* 104.2 (9th Feb. 2023), pp. 1–19. DOI: 10.1099/jgv.0.001827. URL: https://inria.hal.science/hal-04350796.

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

International peer-reviewed conferences

[21] J. Ellert, P. Gawrychowski and G. Gourdel. 'Optimal Square Detection Over General Alphabets'. In: *Proceedings of the 2023 Annual ACM-SIAM Symposium on Discrete Algorithms (SODA)*. SODA - 2023 Annual ACM-SIAM Symposium on Discrete Algorithms. Florence, Italy: Society for Industrial and Applied Mathematics, 16th Feb. 2023, pp. 5220–5242. DOI: 10.1137/1.9781611977554.ch189. URL: https://hal.science/hal-03991054.

[22] P. Gawrychowski, G. Gourdel, T. Starikovskaya and T. A. Steiner. 'Compressed Indexing for Consecutive Occurrences'. In: *34th Annual Symposium on Combinatorial Pattern Matching (CPM 2023)*. CPM 2023 - 34th Annual Symposium on Combinatorial Pattern Matching. Paris, France: Schloss Dagstuhl - Leibniz-Zentrum für Informatik, 2023. DOI: 10.4230/LIPIcs.CPM.2023.12. URL: https://inria.hal.science/hal-04325233.

Conferences without proceedings

[23] C. Berton, G. Coatrieux, D. Lavenier and H. S. 'Dynamic sliding window encoding for data storageon DNA under biological and indexing constraints'. In: EUSIPCO 2023 - 31st European Signal Processing Conference. Helsiinki, Finland, 1st Sept. 2023, pp. 1–5. DOI: 10.23919/EUSIPCO58844.2023.10289957. URL: https://inria.hal.science/hal-04246615.

[24] S. Dubois, B. Linard, M. Zytnicki, C. Lemaitre and T. Faraut. 'Towards an edit distance between pangenome graphs'. In: SeqBIM 2023 - Journées sur les Séquences en Bioinformatique, Informatique et Mathématiques. Lille, France, 2023, pp. 1–2. URL: https://inria.hal.science/hal-04320771.

[25] R. Faure, J.-F. Flot and D. Lavenier. 'HairSplitter: separating strains in metagenome assemblies with long reads'. In: JOBIM 2023 - Journées Ouvertes en Biologie, Informatique et Mathématiques. Plouzané, France, July 2023, pp. 1–8. URL: https://hal.science/hal-04272455.

[26] R. Faure, B. Hilaire and D. Lavenier. 'Mapping-friendly Sequence Reductions to process compressed genomic data'. In: SeqBIM 2023. Lille, France, 2023, pp. 1–1. URL: https://hal.science/hal-04272505.

[27] T. K. M. Truong, R. Faure and R. Andonov. 'Assembling close strains in metagenome assemblies using discrete optimization'. In: BIOINFORMATICS 2024. Rome, Italy, 21st Feb. 2024. URL: https://inria.hal.science/hal-04349675.

Scientific book chapters

[28] C. Guyomar and C. Lemaitre. 'Métagénomique et métatranscriptomique'. In: *Des séquences aux graphesMéthodes et structures discrètes pour la bioinformatique*. ISTE, July 2023, pp. 1–36. URL: https://inria.hal.science/hal-04338891.

Doctoral dissertations and habilitation theses

[29] V. Epain. 'DNA fragment assembly: graph structures and chloroplast genome scaffolding: Comparative analyses, formulations and implementations'. Université de Rennes, 27th Dec. 2023. URL: https://inria.hal.science/tel-04357206.

[30] G. Gourdel. 'Sketch-based approaches to process massive string data'. Université de Rennes, 26th Oct. 2023. URL: https://theses.hal.science/tel-04325241.

[31] L. Robidou. 'Search engine for genomic sequencing data'. Inria, 21st Sept. 2023. URL: https://hal.science/tel-04352906.

Reports & preprints

[32] K. Břinda, L. Lima, S. Pignotti, N. Quinones-Olvera, K. Salikhov, R. Chikhi, G. Kucherov, Z. Iqbal and M. Baym. *Efficient and Robust Search of Microbial Genomes via Phylogenetic Compression*. 16th Apr. 2023. DOI: 10.1101/2023.04.15.536996. URL: https://hal.science/hal-04287842.

[33] S. Dubois, C. Lemaitre, T. Faraut and M. Zytnicki. *Comparaison et visualisation de graphes de pangénomes*. Université de rennes, 8th June 2023, pp. 1–38. URL: https://hal.science/hal-04213245.

[34] V. Epain and R. Andonov. *Global exact optimisations for chloroplast genome multimeric forms scaffolding*. 19th June 2023. URL: https://inria.hal.science/hal-04134429.

[35] P. Gawrychowski, G. Gourdel, T. Starikovskaya and T. A. Steiner. *Compressed Consecutive Pattern Matching*. 20th Oct. 2023. URL: https://hal.science/hal-04251959.

[36] J. Leblanc, O. Boulle, E. Roux, J. Nicolas, D. Lavenier and Y. Audic. *In vitro construction and long read sequencing analysis of a 24 kb long artificial DNA sequence encoding the Universal Declaration of the Rights of Man and of the Citizen*. 26th June 2023. DOI: 10.1101/2023.06.26.546242. URL: https://hal.science/hal-04357386.

[37] T. Lemane, N. Lezzoche, J. Lecubin, E. Pelletier, M. Lescot, R. Chikhi and P. Peterlongo. *kmindex and ORA: indexing and real-time user-friendly queries in terabyte-sized complex genomic datasets*. 4th June 2023. DOI: 10.1101/2023.05.31.543043. URL: https://cea.hal.science/cea-04321497.

[38] C. Moraga, C. Branco, Q. Rougemont, P. Veltsos, P. Jedlička, A. Muyle, M. Hanique, E. Tannier, X. Liu, E. Mendoza-Galindo et al. *The Silene latifolia genome and its giant Y chromosome*. 22nd Sept. 2023. DOI: 10.1101/2023.09.21.558754. URL: https://hal.science/hal-04293712.

[39] O. Sladký, P. Veselý and K. Břinda. *Masked superstrings as a unified framework for textual k-mer set representations*. 3rd Feb. 2023. DOI: 10.1101/2023.02.01.526717. URL: https://inria.hal.science/hal-03970624.

Other scientific publications

[40] R. Faure, J.-F. Flot and D. Lavenier. 'HairSplitter: separating similar strains in metagenome assembly'. In: ISMB/ECCB 2023 - 31st Annual Intelligent Systems For Molecular Biology and the 22nd Annual European Conference on Computational Biology. Lyon, France, 2023, pp. 1–1. URL: https://hal.science/hal-04272480.

[41] N. Maurice. 'Assemblage métagénomique d'écosystèmes complexes avec différentes technologies de séquençage de 3ème génération'. Université de Bordeaux, 12th June 2023, p. 33. URL: https://inria.hal.science/hal-04142837.

[42] S. Robin, A. Bretaudeau and F. Legeai. 'BIPAA, Bioinformatics Platform for the Agroecosystems Arthropods.' In: JOBIM 2023 - Journées Ouvertes en Biologie, Informatique et Mathématiques. Plouzané (Brest Métropole), France, 27th June 2023. URL: https://hal.inrae.fr/hal-04176786.

12.3 Other

Scientific popularization

[43] D. Lavenier, M. Antonini, Y. Rondelez and A. Genot. 'Stocker les données : la piste prometteuse de l'ADN'. In: *Interstices* (1st Feb. 2023). URL: https://inria.hal.science/hal-04073357.

[44] H. Touzet, M. Salson, C. Lemaitre and F. Débarre. 'Le SARS-CoV-2 : une expérience inédite de surveillance génomique mondiale'. In: *Interstices* (10th May 2023). URL: https://inria.hal.science/hal-04155812.

12.4 Cited publications

[45] P. Bille, I. L. Gørtz, M. R. Pedersen and T. A. Steiner. 'Gapped indexing for consecutive occurrences'. In: *arXiv preprint arXiv:2102.02505* (2021).