RESEARCH CENTRE

Inria Center at Rennes University

IN PARTNERSHIP WITH:

CNRS, Université Rennes 1

2022

ACTIVITY REPORT

Project-Team GENSCALE

Scalable, Optimized and Parallel Algorithms for Genomics

IN COLLABORATION WITH: Institut de recherche en informatique et systèmes aléatoires (IRISA)

DOMAIN

Digital Health, Biology and Earth

THEME

Computational Biology

Inria

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 Brinda [INRIA, ISFP]
- Dominique Lavenier [CNRS, Senior Researcher, HDR]
- Claire Lemaitre [INRIA, Researcher, HDR]
- Jacques Nicolas [INRIA, Senior Researcher, HDR]

Faculty Member

- Rumen Andonov [UNIV RENNES , Professor, HDR]

PhD Students

- Kevin Da Silva [Inria, until Mar 2022]
- Clara Delahaye [UNIV RENNES I, (until 31 Dec. 2022)]
- Victor Epain [INRIA]
- Roland Faure [UNIV RENNES I]
- Garance Gourdel [UNIV RENNES I]
- Khodor Hannoush [INRIA]
- Teo Lemane [INRIA, (until 31 Dec. 2022)]
- Meven Mognol [UPMEM, CIFRE, from Mar 2022]
- Lucas Robidou [INRIA]
- Sandra Romain [INRIA]

Technical Staff

- Olivier Boulle [INRIA, Engineer]
- Charles-Adolphe Deltel [INRIA]
- Anne Guichard [INRIA, Engineer]
- Julien Leblanc [CNRS, Engineer]
- Gildas Robine [CNRS, Engineer]

Interns and Apprentices

- Jacky Ame [INRAE , Intern, until Jul 2022]
- Siegfried Dubois [INRIA, until Aug 2022]
- Nathan Merillon [ENS Rennes, Intern, from May 2022 until Jul 2022]
- Emma Redor [INRIA, Intern, from May 2022 until Jul 2022]
- Tam Truong Khac Minh [UNIV RENNES I, Intern, from May 2022 until Jul 2022]

Administrative Assistant

- Marie Le Roic [INRIA]

External Collaborators

- Susete Alves Carvalho [INRAE]
- Fabrice Legeai [INRAE]
- Emeline Roux [UNIV RENNES I, Associate professor]

2 Overall objectives

The main goal of the GenScale project is to develop scalable methods, tools, and software 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: Parallelism
- Axis 3: Sequence analyses algorithms
- Axis 4: Applications

3 Research program

3.1 Axis 1: Data Structures

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. 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 [7].

3.2 Axis 2: 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 implementation 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 [9], to detect structural variants using local NGS assembly approaches [8] 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 [16].

- **Storing information on DNA molecules** DNA molecule can be seen as promising support for information storage. This can be achieved by encoding information into DNA alphabet, including error correction codes, 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.

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. This 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 feedback.

4 Application domains

4.1 Introduction

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.

4.2 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.

4.3 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.

4.4 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 Massive indexing of genomic data

The kmtricks tool, published in bioinformatics advance [16] and presented in this report section 8.1.1, represents an important step towards massive indexing of large genomic databases. This is the first tool able to index dozens of terabytes of raw sequencing data, with a final index size of ≈ 10% of the input compressed data. It allows hundreds of queries to be answered in a few tens of minutes.

More recently, we derived from kmtricks another tool, kmindex. This novel tool, unpublished yet, generates bigger indexes (approximately 10% bigger on metagenomics data). Its advantages are: (1) indexing time is roughly two times faster that kmtricks (2) query time takes few milliseconds instead of minutes, being thus ≈1700 times faster.

Using kmindex, ≈ 37 terabytes of raw compressed metagenomics data from the Tara ocean project were indexed in 7 days. Using previous approaches, the same task would need several months to achieve. The associated search engine is currently in deployment on the OGA server.

6.2 Organisation of JOBIM 2022

GenScale organized JOBIM 2022, the 23th edition of the French conference on computational biology, from July 5th to 8th at University of Rennes (web site). This event was a success and gathered more participants than previous editions, with 536 participants during the 4 days. We invited 6 international keynote speakers and received 300 submissions. The program was composed of 6 keynote presentation, 42 accepted oral contributions, 17 demos and 240 posters. Different research networks took advantage of the JOBIM conference to meet and exchange, notably around 5 thematic workshops.

The organization of this conference mobilized the whole team at the time of the conference but also upstream. In particular, two members of the team were co-chairs of the organizing and program committees.

7 New software and platforms

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 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

7.1.3 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 overestimaed calls. fimpera reduces their false positive rate by two order of magnitude while reducing the overestimations, whithout introducing false negative and by speeding up queries.

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

Contact: Lucas Robidou

Participants: Lucas Robidou, Pierre Peterlongo

7.1.4 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

7.1.5 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

7.1.6 QuickDeconvolution

Keywords: High throughput sequencing, Genomics

Functional Description: QuickDeconvolution deconvolutes a set of linked reads: QuickDeconvolution takes as input a linked reads dataset and adds an extension (-1, -2, -3...) to the barcodes, such that two reads with the same barcode and the same extension comes from the same genomic region.

Release Contributions: This new versions implement a series of improvements that have been made in order to publish the paper.

URL: http://github.com/RolandFaure/QuickDeconvolution

Contact: Roland Faure

Participants: Roland Faure, Dominique Lavenier

7.1.7 GraphUnzip

Keywords: Genome assembly, Genome assembling, Haplotyping

Functional Description: GraphUnzip untangles assembly graphs: GraphUnzip takes two input: 1) An assembly graph in GFA fromat, from an assembler 2) Data that can help untangling the graph: Hi-C, long reads or linked reads.

GraphUnzip returns an untangled assembly graph, improving significantly the contiguity of the input assembly.

Release Contributions: Brand new Hi-C algorithm now fit for publication. Great increase in accuracy and performance.

URL: http://github.com/nadegeGuiglielmoni/GraphUnzip

Contact: Roland Faure

Participants: Roland Faure, Jean-François Flot, Nadège Guiglielmoni

Partner: Université libre de Bruxelles

7.1.8 SeqFaiLR

Keywords: Long reads, Sequencing error, Sequence alignment

Functional Description: SeqFaiLR analyses Nanopore long reads sequencing error profiles. The algorithms have been designed for Nanopore data, but can be applied for other long read data. From raw reads and reference genomes, these scripts perform alignment and compute several analysis (low-complexity regions sequencing accuracy, GC bias, links between error rates and quality scores, and so on).

URL: https://github.com/cdelahaye/SeqFaiLR

Contact: Clara Delahaye

Participants: Jacques Nicolas, Clara Delahaye

7.1.9 ORI

Name: Oxford nanopore Reads Identification

Keywords: Bioinformatics, Bloom filter, Spaced seeds, Long reads, ASP - Answer Set Programming, Bacterial strains

Functional Description: ORI (Oxford nanopore Reads Identification) is a software using long nanopore reads to identify bacteria present in a sample at the strain level. There are two sub-parts in ORI: (1) the creation of the index containing the reference genomes of the interest species and (2) the query of this index with long reads from Nanopore sequencing in order to identify the strain(s).

URL: https://github.com/gsiekaniec/ORI

Contact: Jacques Nicolas

Participants: Gregoire Siekaniec, Teo Lemane, Jacques Nicolas, Emeline Roux

7.1.10 Wisp

Name: A Python application for bacterial families identification from long reads.

Keywords: Nucleic Acids, Machine learning, Bioinformatics, Biodiversity, Genomic sequence, Omic data, Bacterial strains

Scientific Description: Genomic and metagenomic data are flowing into microbiology thanks to advances in sequencing. In particular, we consider here ONT Minion long read data, that have a relatively high error rate. In this context, this work addresses a key problem, binning, which consists of grouping sequenced reads into taxonomically coherent sets. We have learned genomic signatures on two databases of microbial genomes and for various taxonomic levels, relying on a model of regression trees boosting, thanks to the XGBoost library. We used as attributes the frequencies of small k-mers (in range 4-6) on 10kb fragments sampled along the genomes. Each level of the taxonomy (until the family level) is predicted assuming the previous level is known. The prediction was made at the scale of single reads and groups of 400 reads, with a preprocessing step discarding low signifiance fragments. Overall, the level of accuracy achieved is very satisfactory, with known occasional issues such as for the Fusobacteria phylum and the Deltaproteobacteria group. Apart from the domain level (bacteria or archae), the most coherent taxonomic level seems to be that of the order. We designed a software suite, Wisp, covering database creation, error processing, creation of reports, quality analysis of predictions, using a user-friendly parameters interface.

Functional Description: This tool is requiring some reference genomes, which it will index, to create a XGBoost model. Genomic fasta (.fna) files are the prefferd input style as of now. One can download custom genome dataset with NCBI accession numbers to create its own specific dataset and increase even more classifier accuracy.

Release Contributions: Initial version

Contact: Jacques Nicolas

Participants: Jacques Nicolas, Siegfried Dubois

7.1.11 BlockDTW

Name: Block Dynamic Time Warping

Keywords: Alignment, Algorithm, Distance

Functional Description: This software provides implementations of an O(nM+mN) and an O(nmk) time algorithm to compute pattern matching for DTW distance. We divide the dynamic programming matrix into blocks and improve the computation inside those, for further details see https://hal.archives-ouvertes.fr/hal-03763091/

URL: https://github.com/fnareoh/DTW

Contact: Garance Gourdel

Participants: Garance Gourdel, Anne Driemel, Pierre Peterlongo, Tatiana Starikovskaya

7.1.12 PyRevSymG

Name: Python Reverse Symmetric Graph

Keywords: Directed graphs, Graph algorithmics, DNA sequencing

Functional Description: 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

7.1.13 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

7.1.14 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

8 New results

8.1 Indexing data structures

8.1.1 Represent large sets of sequencing data with kmer matrices or bloom filters

Participants: Pierre Peterlongo, Téo Lemane.

When indexing large collections of short-read sequencing data, a common operation that has now been implemented in several tools (Sequence Bloom Trees [51], Mantis [50] BIGSI [49] and variants) is to construct a collection of Bloom filters, one per sample. Each Bloom filter is used to represent a set of kmers. kmers whose abundance is lower than a hard threshold are discarded. This representation approximates the desired set of all the non-erroneous kmers present in the sample. It has the precious advantage to index complete read sets composed of hundreds of millions reads with a few GB of memory or disk. However, this approximation is imperfect, especially in the case of metagenomics data. Erroneous but abundant kmers are wrongly included, and non-erroneous but low-abundant ones are wrongly discarded. Additionally, existing tools able to generate matrices of counted kmers or collections of bloom filters have important running time.

In this context, we proposed *kmtricks* [16], a novel approach for generating Bloom filters from terabase-sized collections of sequencing data. Our main contributions are 1/ an efficient method for jointly counting kmers across multiple samples, including a streamlined Bloom filter construction by directly counting, partitioning, and sorting hashes instead of kmers, which is approximately four times faster than state-of-the-art tools; 2/ a novel technique that takes advantage of joint counting to preserve low-abundant kmers present in several samples, improving the recovery of non-erroneous kmers. Our experiments highlight that this technique preserves around 8x more valid kmers than the usual yet crude filtering of low-abundance kmers in a large metagenomics dataset.

Using our workflow, we performed for the first time a massive-scale joint kmer counting and Bloom filter construction of a 6.5 terabases metagenomics collection, in under 50 GB of memory and 38 hours, which is at least 3.8 times faster than the next best alternative.

8.1.2 kmer-based genome wide association studies

Participants: Pierre Peterlongo, Téo Lemane.

Genome Wide Association Studies (GWAS) elucidate links between genotypes and phenotypes. Recent studies point out the interest of conducting such experiments using kmers as the base signal instead of single-nucleotide polymorphisms. We proposed a tool, called kmdiff, that performs differential kmer analyses on large sequencing cohorts in an order of magnitude less time and memory than previously possible [15].

8.1.3 Improvement of Approximate Membership Query data-structures with counts

Participants: Pierre Peterlongo, Lucas Robidou.

Approximate membership query data structures (AMQ) such as Cuckoo filters or Bloom filters are widely used for representing and indexing large sets of elements. AMQ can be generalized for additionally counting indexed elements, they are then called “counting AMQ”. This is for instance the case of the “counting Bloom filters”. However, counting AMQs suffer from false positive and overestimated calls.

We proposed a novel computation method, called fimpera, consisting of a simple strategy for reducing the false-positive rate of any AMQ indexing all kmers from a set of sequences, along with their abundance

information. This method decreases the false-positive rate of a counting Bloom filter by an order of magnitude while reducing the number of overestimated calls, as well as lowering the average difference between the overestimated calls and the ground truth. In addition, it slightly decreases the query run time. The fimpera method does not require any modification of the original counting Bloom filter, it does not generate false-negative calls, and it causes no memory overhead. The unique drawback is that fimpera yields a new kind of false positives and overestimated calls. However their amount is negligible. As a side note, for the algorithmic needs of the method, we also propose a novel generic algorithm for finding minimal values of a sliding window over a vector of x integers in $O(x)$ time with zero memory allocation [41].

8.1.4 Standardized and compact disk representation of sets of k-mers

Participants: Pierre Peterlongo, Téo Lemane.

Bioinformatics applications increasingly rely on ad hoc disk storage of kmer sets, e.g. for de Bruijn graphs or alignment indexes. Here, we introduce the Kmer File Format (KFF) as a general lossless framework for storing and manipulating kmer sets, realizing space savings of 3–5× compared to other formats, and bringing interoperability across tools. [11]

8.2 Algorithms for genome assembly and variant detection

8.2.1 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, which uses sequence graphs instead of linear sequences to represent the SVs.

In our method, we build a variation graph from a reference genome and a given set of SVs. The SV breakpoints are extracted and sorted. The genome sequence is then split at each breakpoint into non-overlapping fragments. Each fragment becomes a node in the graph, and edges are added between nodes to form the reference path of the genome as well as the alternative path for each SV. Additional nodes are added for insertions. The long reads are then mapped on the variation graph and the resulting alignments that overlap an edge (breakpoint) in the graph are used to estimate the most likely genotype for each SV.

Running SVJedi-graph on simulated sets of close deletions showed that the use of a variation graph was able to restore the genotyping quality on close and overlapping SVs. For instance, with a simulated set of deletions that had another deletion 0 to 50 bp apart, SVJedi-graph was able to genotype 99.6% of the deletions with an accuracy of 98.5%, compared to a genotyping rate of 78.9% and an accuracy of 97.3% with SVJedi on the same dataset. We tested our method on a "gold standard" datasets of Genome In A Bottle (Tier 1 SVs of human individual HG002), and obtained higher genotyping rates than SVJedi and a higher genotyping accuracy than other state of the art tools [45].

8.2.2 Deconvolution of linked-read sequencing data

Participants: Roland Faure, Dominique Lavenier.

Linked-read technologies, such as the 10X chromium system, use microfluidics to tag multiple short reads from the same long fragment (50–200 kb) with a small sequence, called a barcode. They are inexpensive and easy to prepare, combining the accuracy of short-read sequencing with the long-range information of barcodes. The same barcode can be used for several different fragments, which complicates the analyses. We have developed QuickDeconvolution (QD), a fast software for deconvolving a set of reads sharing a barcode, i.e. separating the reads from the different fragments. QD only takes sequencing data as input, without the need for a reference genome. We showed that QD outperforms existing software in terms of accuracy, speed and scalability, making it capable of deconvolving previously inaccessible data sets [12].

8.2.3 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, there is a lack of local assembly tools for linked-read data.

We present MTG-Link, 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 [8] 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* [40].

8.2.4 Assembling unknown numbers of haplotypes

Participants: Roland Faure, Dominique Lavenier.

We are currently designing a software to provide phased assemblies from draft, purged assemblies. It implements a new method that takes as input a contig and the set of (high-error-rate) sequencing reads used to build it. Then it finds out whether the contig is actually a mix of several haplotypes, and if so outputs the haplotype-specific versions of this contig. To do so, SNPs are called rudimentarily allowing recurring pattern detection of variants among the reads. Unlike previously existing techniques, the method does not need as an input the number of expected haplotypes, as each recurring pattern of variants results in one group of reads. This makes the method especially useful to assemble polyploid species (common in plants and fishes), metagenomic samples, and even repeated regions [27] [44].

8.2.5 Haplotype phasing of long reads for polyploid species

Participants: Clara Delahaye, Jacques Nicolas.

We are working on a binning problem, assigning the long reads of a sample to their native haplotype. While existing approaches rely on the use of a reference genome and known variants, we propose a method relying on long reads only, suited for the phasing of polyploid organisms. In order to compensate for the absence of reference genome, independent sub-problem instances are built from a restrained set of the longest reads acting as a pseudo-reference. The longest reads are phased and then used as anchors to map remaining reads, which are phased by maximizing their consistency with the phased anchors. By reasoning on the set of possible solutions, and integrating user preferences, it is possible to increase the robustness of results. We applied the phasing algorithm on a diploid, *A. vaga* (showing ancient tetraploid state) long read dataset for which a confident phased reference genome is available to evaluate the results. We also introduced an haplotig graph, which enables to explicitly point the regions of identity between haplotypes, as well as their differences.

This work has led to the PhD defense of Clara Delahaye at the end of this year [33].

8.2.6 Efficiently storing DNA fragments' succession relationships in a graph

Participants: Victor Epain, Rumen Andonov.

Assembling DNA fragments based on their overlaps remains the main assembly paradigm with long DNA fragments sequencing technologies, independently of the aim to resolve only one or several haplotypes. Since an overlap can be seen as a succession relationship between two oriented fragments, the directed graph structure has emerged as an appropriate data structure for handling overlaps. However, this graph paradigm does not appear to take advantage of the reverse symmetry of the orientated fragments and their overlaps, which is a result of blind DNA double-strand sequencing. Thus, the bi-directed graph paradigm was introduced in 1995 towards reducing the graph size by handling the reverse symmetry, and becomes since then the main graph paradigm used in assembly/scaffolding methods. Nevertheless, these two graph paradigms have never been contrasted before, and no implementations have been described. We present suitable data structures that are theoretically compared in terms of time and memory consumption in the context of the design of some basic graph algorithms. We also show that each one of the paradigms can be switched to another by slightly modifying their data structures.

These results are described in a submitted version for RECOMB2022 conference [38]. They have been presented at DBS2022 workshop at Düsseldorf. An extended version can be found in [36].

One of the described graph implementations is available on a public released Python3 package 7.1.12

8.2.7 Optimal inverted repeats scaffolding for chloroplast genomes

Participants: Victor Epain, Rumen Andonov, Dominique Lavenier.

Here we describe a novel assembly approach for chloroplast genomes. It contains two modular steps. In the first step, based on the hypothesis that chloroplasts genomes are over-represented compared to the nuclear genome in the plant's cell, we assemble contigs with a De Bruijn graph based approach using short reads with a high k-mer coverage. Connections between oriented contigs are also provided here. The second step determines the order and the orientation of the contigs (scaffolding). Taking advantage of the knowledge that chloroplast genomes possess well studied circular structure, we develop a particular formulation of the scaffolding problem, called Nested Inverted Fragments Scaffolding. It aims at assembling highly conserved inverted repeats. We formulate it as an optimisation problem and we prove that it is NP-Complete. To solve the problem we propose and implement an integer linear programming formulation. We evaluate our method on a set of real instances (a benchmark of 42 chloroplast genomes) and show that it obtains notable achievements with respect to the quality of the results. To further estimate the performance of our scaffolding module, we test it on huge artificially created instances. The results demonstrate an excellent behaviour of our integer formulation as even very large instances have been solved at the first Branch and Bounds node.

These works have been presented at the ROADEF2022 conference [26] and at the JOBIM2022 conference [22]. These results are described in a submitted version for ISCO2022 symposium [37] and in a submitted version for WABI2023 conference [39].

8.3 Information storage on DNA molecules

8.3.1 Storing the declaration of human rights on a single DNA molecule

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

Today, the community consensus to store information on DNA is to use short single strand DNA (ssDNA) molecules. This approach has some limitations: encoding constraint, DNA stability, recovering DNA, sequencing technology, etc. To overcome them, we chose to store information on long double-strand DNA (dsDNA) molecules. Our demonstration consists in storing the first articles of the declaration of human rights (4.2 KByte text document) on a single DNA molecule [43].

Our approach is based on an ordered assembly of short oligonucleotides. The document is first split into small DNA fragments whose length are compatible with DNA synthetizers (about 60 nucleotides). A first assembly concatenates pools of 10 oligonucleotides to form double strand DNA molecules of approximatively 600 bp. A new round of ordered assembly takes these 600 bp molecules to build 6 kbp molecules. The last round assembles 5 of these molecules to form the final long DNA molecules supporting the full text.

To be able to build such long DNA molecules, we have developed a specific and systematic biotechnology protocol which, in interaction with our experimental platform (see next paragraph) should lead to a complete automation of the DNA writing process.

8.3.2 Experimental DNA storage platform

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

The dnarXiv projects aims to explore various strategies for DNA storage. We have designed an experimental DNA storage platform allowing both to conduct real and/or in-silico experiments. The platform includes different modules generally used in the write/read DNA storage process: encoding, synthesis, molecule design, storage, molecule selection, genomic data processing, decoding. This is a flexible environment for testing various approaches simply by substituting new modules to the existing ones [25].

8.3.3 DNA data storage security

Participants: Dominique Lavenier.

In this work, we are interested in securing archived data. DNA storage being a new technology, there is an opportunity to integrate this critical aspect at the biological level, contrarily to what has been done for electronical storage means. In fact, information must be secured at every step of the DNA data storage chain. Data integrity and confidentiality are among the main issues with threats like data modification (e.g. writing of new data) or the theft of the DNA storage support by an attacker. Herein, we propose a solution for writing encrypted data onto synthetic DNA molecules considering DNA synthesis and the error-correction code constraints. Indeed, DNA sequences should conform to structural constraints dictated by this biological process and sequencing [24] [42].

8.3.4 Exploring DNA synthesis and sequencing semiconductor technologies

Participants: Dominique Lavenier.

Within the dnarXiv project, we explored the different ways of performing synthesis and sequencing steps and more specifically how the reading and writing processes can be implemented on semiconductor devices. [28]

8.4 Processing-in-Memory

Participants: Karel Brinda, Charles Deltel, Dominique Lavenier, Meven Mognol, Gildas Robine.

All current computing platforms are designed following the von Neumann architecture principles, originated in the 1940s, that separate computing units from memory. 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.

New memory components based on PIM principles have been developed by the UPMEM company, a young startup created in 2015. The company has designed an innovative DRAM Processing Unit (DPU), a RISC processor integrated directly in the memory chip, on the DRAM die. An UPMEM PIM server counts no less than 2560 DPUs for 160 GB of PIM memory and 256 GB of legacy memory. First experiments on the UPMEM PIM server have demonstrated that an average speed-up of X20 can generally be obtained on various time-consuming tasks of NGS pipelines compared to standard multicore platforms [29].

In 2022, we specifically worked on the three following algorithms:

1. Sequence alignment with the KSW2 software
2. Bacterial genome comparison
3. Protein sequence alignment

Sequence Alignment. The aim is to compute consensus sequence from long reads. It implies to make many pairwise comparisons based on dynamic programming algorithms. KSW2 is used for that purpose. We are currently implementing a processing-in-memory parallel strategy of KSW2 to optimize the full process. The last measurements show a speed-up ranging from 5 to 10 compared to an optimized openMP implementation.

Bacterial genome comparison. The goal is to compare large sets of bacterial genomes to estimate their similarities (as DASHING2). Genome sketches are first computed, and distances are computed based on these sketches. We face a massive parallelism that efficiently exploit on the UPMEM server. Speed-up around 20 is expected. Experimentation are done on a set of 661 000 bacterial genomes.

Protein sequence alignment. We are currently implementing a blast-like algorithm to scan large protein databanks. The protein bank is split over the Processing-in-Memory components. The query is broadcasted to all DRAM processing units (DPU) which send back alignments to the host processor. The parallelization of the algorithm is achieved and performance measurements will start in January 2023.

8.5 Benchmarks and Reviews

8.5.1 Evaluation of metagenomic software: the second round of CAMI challenges

Participants: Claire Lemaitre, Pierre Peterlongo.

Evaluating metagenomic software is key for optimizing metagenome interpretation and focus of the Initiative for the Critical Assessment of Metagenome Interpretation (CAMI). The CAMI II challenge engaged the community to assess methods on realistic and complex datasets with long- and short-read sequences, created computationally from around 1,700 new and known genomes, as well as 600 new plasmids and viruses. Here 5,002 results by 76 program versions were analyzed. Substantial improvements were seen in assembly, some due to long-read data. Related strains still were challenging for assembly and genome recovery through binning. Runtime and memory usage analyses identified efficient programs, including top performers with other metrics. The results identify challenges and guide researchers in selecting methods for analyses [18].

GenScale team members have participated in this competition: we runned our genome assembly software [4] on the CAMI data and provided the results and full pipelines to the evaluation team.

8.5.2 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 [31].

8.6 Theoretical studies

8.6.1 Pattern matching under DTW distance

Participants: Garance Gourdel, Pierre Peterlongo.

We considered the problem of pattern matching under the dynamic time warping (DTW) distance motivated by potential applications in the analysis of biological data produced by the third generation sequencing. To measure the DTW distance between two strings, we must “warp” them, that is, double some letters in the strings to obtain two equal-length strings, and then sum the distances between the letters in the corresponding positions. When the distances between letters are integers, we show that for a pattern P with m runs (a run being a maximal set of consecutive letters) and a text T with n runs there is an $O(kmn)$-time algorithm that computes all locations where the DTW distance from P to T is at most k [23].

8.6.2 Streaming Regular Expression Membership and Pattern Matching

Participants: Garance Gourdel.

A regular expression R is a formalism for compactly describing a set of strings, built recursively from single characters using three operators: concatenation, union, and Kleene star. In this paper we study membership and pattern matching of regular expressions in the streaming setting (where the pattern can be preprocessed, then the characters of the text are seen one at a time, and all space must be accounted for). In general, we cannot hope for a streaming algorithm with space complexity smaller than the length of R for either variant of regular expression search. The main contribution of this paper is that we identify the number of unions and Kleene stars, denoted by d, as the parameter that allows for an efficient streaming algorithm. We design general randomised Monte Carlo algorithms for both problems that use $O(d^3 \text{ polylog } n)$ space in the streaming setting [30].

8.7 Bioinformatics Analysis

8.7.1 Comparing seawater metagenomes from the Tara ocean project

Participants: Claire Lemaitre, Pierre Peterlongo.

Biogeographical studies have traditionally focused on readily visible organisms, but recent technological advances are enabling analyses of the large-scale distribution of microscopic organisms, whose biogeographical patterns have long been debated. Here, we assess global plankton biogeography and its relation to the biological, chemical and physical context of the ocean (the 'seascape') by analyzing 24 terabases of metagenomic sequence data and 739 million metabarcodes from the Tara Oceans expedition in light of environmental data and simulated ocean current transport. We show tat, in addition to significant local heterogeneity, viral, prokaryotic and eukaryotic plankton communities all display near steady-state, large-scale, size-dependent biogeographical patterns. Correlation analyses between plankton transport time and metagenomic or environmental dissimilarity reveal the existence of basin-scale biological and environmental continua emerging within the main current systems. Across oceans, there is a measurable, continuous change within communities and environmental factors up to an average of 1.5 years of travel time. Finally, modulation of plankton communities during transport varies with organismal size, such that the distribution of smaller plankton best matches Longhurst biogeochemical provinces, whereas larger plankton group into larger provinces [19].

GenScale team members have participated to the developement of algorithms that enable such large scale sequencing data comparisons, and they provided their expertise regarding the results and their analyses.

8.7.2 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. In particular, we participated in the genome assembly and analyses of some major agricultural pests or their natural ennemies such as parasitoids. In most cases, the genomes and their annotations were hosted in the BIPAA information system, allowing collaborative curation of various set of genes and leading to novel biological findings [17, 10, 13, 21].

8.7.3 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].

8.7.4 Genomics of a lactic acid bacterium of industrial and health interest

Participants: Jacques Nicolas, Emeline Roux.

Streptococcus thermophilus is a bacterium widely used in the dairy industry as well as in many traditional fermented products. In addition, *S. thermophilus* exhibits functionalities favorable to Human health. We investigate the main health-promoting properties of *S. thermophilus* and study their intra-species diversity within a collection of representative strains (around 80 genome sequences of strains, 30 of which were sequenced and assembled during Gregoire Siekaniec's thesis). Some functions are widely conserved among isolates (e.g., folate production, degradation of lactose) suggesting their central role for the species, while others (e.g., catabolism of galactose, production of bioactive peptides) are strain-specific. A better understanding of the health-promoting properties and the genomic and genetic diversity within *S. thermophilus* species could facilitate the selection and development of fermented products with health-promoting properties [20, 46].

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 European initiatives

10.1.1 H2020 projects

IGNITE ITN

Participants: Anne Guichard, Pierre Peterlongo.

IGNITE project on cordis.europa.eu

Title: Comparative genomics of non-model invertebrates

Duration: From January 1, 2018 to June 30, 2022

Partners:

- INSTITUT NATIONAL DE RECHERCHE EN INFORMATIQUE ET AUTOMATIQUE (INRIA), France
- LUDWIG-MAXIMILIANS-UNIVERSITAET MUENCHEN (LMU MUENCHEN), Germany
- NATIONAL UNIVERSITY OF IRELAND GALWAY (NUI GALWAY), Ireland
- FACULTY OF SCIENCE UNIVERSITY OF ZAGREB (FACULTY OFSCIENCE UNIVERSITY OF ZAGREB), Croatia
- EUROPEAN MOLECULAR BIOLOGY LABORATORY (EMBL), Germany
- GENERALDIREKTION DER STAATLICHE NATURWISSENSCHAFTLICHEN SAMMLUNGEN BAYERNS (SNSB), Germany
- HITS GGMBH (HITS), Germany
- UNIVERSITE LIBRE DE BRUXELLES (ULB), Belgium
- BAYERISCHE AKADEMIE DER WISSENSCHAFTEN (BADW), Germany
- Era7 Information Technologies SL (Era7), Spain
- INSTITUT PASTEUR (IP), France
- UNIVERSITY OF BRISTOL, United Kingdom
- Fourmy (Alphabiotoxine), Belgium
- CENTRO INTERDISCIPLINAR DE INVESTIGACAO MARINHA E AMBIENTAL (CENTRO INTERDISCIPLINAR DE INVESTIGACAO MARINHA E AMBIENTAL), Portugal
- PENSOFT PUBLISHERS (PENSOFT), Bulgaria
- Board of the Queensland Museum (Queensland Museum Network), Australia
- INSTITUT NATIONAL DE RECHERCHE POUR L'AGRICULTURE, L'ALIMENTATION ET L'ENVIRONNEMENT (INRAE), France
- UNIVERSITY COLLEGE LONDON, United Kingdom
- UNIVERSITETET I BERGEN (UiB), Norway

Inria contact: Pierre Peterlongo

Coordinator: Gert Wörheide, Ludwig-Maximilians-Universität München, Germany

Summary: Invertebrates, i.e., animals without a backbone, represent 95% of animal diversity on earth but are a surprisingly underexplored reservoir of genetic resources. The content and architecture of their genomes remain poorly known or understood, but such knowledge is needed to fully appreciate their evolutionary, ecological and socio–economic importance, as well as to leverage the benefits they can provide to human well-being, for example as a source for novel drugs and biomimetic materials. Europe is home to significant world-leading expertise in invertebrate genomics but research and training efforts are as yet uncoordinated. IGNITE will bundle this European excellence to train a new generation of scientists skilled in all aspects of invertebrate genomics. We will considerably enhance animal genome knowledge by generating and analysing novel data from undersampled invertebrate lineages and developing innovative new tools for high-quality genome assembly and analysis. Well-trained genomicists are in increasing demand in universities, research institutions, as well as in software, biomedical, and pharmaceutical companies. Through their excellent interdisciplinary and intersectoral training spanning from biology and geobiology to bioinformatics and computer science, our graduates will be in a prime position to take up leadership roles in both academia and industry in order to drive the complex changes needed to advance sustainability of our knowledge-based society and economy.

ALPACA ITN

Participants: Khodor Hannoush, Pierre Peterlongo.

ALPACA project on cordis.europa.eu

Title: ALgorithms for PAngenome Computational Analysis

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

Partners:

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

Inria contact: Pierre Peterlongo

Coordinator: Alexander Schönhuth, Univ.Bielefeld, Germany

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 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.

BioPIM Project

Participants: Dominique Lavenier, Meven Mognol.

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, Bliken 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.

10.1.2 Other european programs/initiatives

Université Libre de Bruxelles, Belgique. Within the framework of the PhD co-supervision of Roland Faure, we work an genome assembly strategies to extract haplotypes of polyploid genomes.

10.2 National initiatives

10.2.1 PEPR

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

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

Coordinators: Marc Antonini

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

Partners: I3S, LabSTIC, GenScale Irisa/Inria, IPMC, Eurocom

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 subgroup 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.

10.2.2 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. Després (Laboratoire d'écologie alpine (LECA), UMR CNRS 5553, Grenoble)
- Duration: 42 months (Jan. 2021 – Dec. 2024)
- Partners: LECA, UMR CNRS 5553, Grenoble; CEFE, UMR CNRS 5175, Montpellier; Genscale Inria/IRISA Rennes.
- Description: The Divalps project aims at better understanding how populations adapt to changes in their environment, and in particular climatic and biotic changes with altitude. Here, we focus on a complex of butterfly species distributed along the alpine altitudinal gradient. We will analyse the genomes of butterflies in contact zones to identify introgressions and rearrangements between taxa.

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

Project GenoPIM. Processing-in-Memory for Genomics

Participants: Charles Deltel, 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/

10.2.3 Inria Exploratory Action

DNA-based data storage system

Participants: Olivier Boulle, Charles Deltel, Dominique Lavenier, Jacques Nicolas.

- Coordinator : Dominique Lavenier
- Duration : 24 months (Oct. 2020, Sep. 2022)
- Description: The goal of this Inria's Exploratory Action is to develop a large-scale multi-user DNA-based data storage system that is reliable, secure, efficient, affordable and with random access. For this, two key promising biotechnologies are considered: enzymatic DNA synthesis and DNA nanopore sequencing. In this action, the focus is made on the design of a prototype platform allowing in-silico and real experimentations to be done. It is a complementary work with the dnarXiv project.

10.3 Regional initiatives

10.3.1 Labex Cominlabs

Project dnarXiv: Archiving information on DNA molecules

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

- Coordinator : Dominique Lavenier
- Duration : 39 months (Oct. 2020, Dec. 2023)
- 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. We aim 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.
- website: https://project.inria.fr/dnarxiv/

11 Dissemination

11.1 Promoting scientific activities

Participants: Rumen Andonov, Karel Brinda, Victor Epain, Roland Faure, Garance Gourdel, Dominique Lavenier, Fabrice Legeai, Claire Lemaitre, Jaques Nicolas, Pierre Perterlongo.

11.1.1 Scientific events: organisation

General chair

- JOBIM 2022: French symposium of Bioinformatics (500 participants, 4 days) [F. Legeai]

Member of the organizing committees

- JOBIM 2022: French symposium of Bioinformatics (500 participants, 4 days) [the whole team]

11.1.2 Scientific events: selection

Chair of conference program committees

- seqBIM2022: national meeting of the sequence algorithms GT seqBIM, Bordeaux, Nov 2022 (45 participants, 2 days) [C. Lemaitre]
- JOBIM 2022: French symposium of Bioinformatics (6 keynotes, 5 mini-symposia, 300 submissions) [C. Lemaitre]

Member of conference program committees

- ISMB 2022: 30th Conference on Intelligent Systems for Molecular Biology, Madison, Wisconsin, USA, 2022 [D. Lavenier]
- ROADEF 2022: 23th Symposium of the French Society on Operatioonal research [R. Andonov, V. Epain]
- JOBIM 2022: French symposium of Bioinformatics, Rennes, France, 2022 [P. Peterlongo]

Reviewer

- CPM 2022 33rd Annual Symposium on Combinatorial Pattern Matching [G. Gourdel, P. Peterlongo]
- SWAT 2022 18th Scandinavian Symposium and Workshops on Algorithm Theory [G. Gourdel]
- RECOMB 2022 26th Annual International Conference on Research in Computational Molecular Biology [P. Peterlongo]
- SPIRE 2022 29th edition of the annual Symposium on String Processing and Information Retrieval [P. Peterlongo]
- IPDPS 2022 36th IEEE International Parallel & Distributed Processing Symposium [P. Peterlongo]

11.1.3 Journal

Member of the editorial boards

- Insects [F. Legeai]

Reviewer - reviewing activities

- NAR Genomics and Bioinformatics [K. Brinda, R. Faure]
- Microbial Genomics [K. Brinda, G. Gourdel]

11.1.4 Invited talks

- C. Lemaitre, "Methodological challenges of Structural Variation characterization and the particular case of insertions", keynote speaker at the meeting reads2genpop : From sequence reads to genomes and populations, Paris, Sept. 2022.
- C. Lemaitre, "Finding structural variants with sequencing data: the difficult case of insertions", Keynote speaker at the annual meeting of GDR AIEM and Alphy working group (GDR BIM), Rennes, March 2022.
- D. Lavenier, "DNA Storage", IDA 2022, International Symposium on Intelligent Data Analysis, Rennes, Juillet 2022
- P. Peterlongo, "Swim in the data tsunami", JOBIM 2022, Rennes, Juillet 2022, keynote speaker
- K. Brinda, "The tree of life enables efficient and robust compression and search of microbes", JOBIM 2022 minisymposium, Rennes, July 2022.

11.1.5 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]

11.1.6 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]

11.1.7 Research administration

- Corresponding member of COERLE (Inria Operational Committee for the assessment of Legal and Ethical risks). Participation to the ethical group of IFB (French Elixir node, Institut Français de Bioinformatique) [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) [J. Nicolas]
- Chair of the committee in charge of all the temporary recruitments ("Commission Personnel") at Inria Rennes-Bretagne Atlantique and IRISA [D. Lavenier]

11.2 Teaching - Supervision - Juries

Participants: Rumen Andonov, Karel Brinda, Roland Faure, Khodor Hannoush, Dominique Lavenier, Fabrice Legeai, Claire Lemaitre, Jaques Nicolas, Pierre Perterlongo, Lucas Robidou, Sandra Romain, Emeline Roux.

11.2.1 Teaching administration

- In charge of the master's degree "Nutrition Sciences des Aliments" (NSA) of University of Rennes 1 (45 students) [E. Roux]

11.2.2 Teaching

- Licence : R. Andonov, Models and Algorithms in Graphs, 100h, L3, Univ. Rennes 1, France.
- Licence : E. Roux, biochemistry, 90h, L1 and L3, Univ. Rennes 1, France.
- License: L. Robidou, K. Hannoush, Principles of Computer Systems, 48h, L1, Univ. Rennes 1, France.
- License: L. Robidou, Outils bureautiques pour le statisticien, 6h, L1, Ensai, France
- Master : R. Andonov, Operations Research (OR), 82h, M1 Miage, Univ. Rennes 1, France.
- Master : R. Andonov, Optimisation Techniques in Bioinformatics, 18h, M2, Univ. Rennes 1, France.
- Master : C. Lemaitre, P. Peterlongo, S. Romain, Algorithms on Sequences, 52h, M2, Univ. Rennes 1, France.
- Master : C. Lemaitre, S. Romain, Bioinformatics of Sequences, 40h, M1, Univ. Rennes 1, France.
- Master : P. Peterlongo, Experimental Bioinformactics, 24h, M1, ENS Rennes, France.
- Master : E. Roux, biochemistry, 120h, M1 and M2, Univ. Rennes 1, France.
- Aggreg: D. Lavenier, Computer Architecture, 20h, ENS Rennes

11.2.3 Supervision

- PhD: C. Delahaye, haplotype phasing from long reads with ASP: a flexible optimization approach [33], defended: 15/12/2022, J. Nicolas.
- PhD: T. Lemane, unbiased detection of neurodegenerative structural variants using k-mer matrices [34], defended: 16/12/2022, P. Peterlongo.
- PhD: K. Da Silva, Identification and quantification of microbial strains in metagenomic samples using variation graphs [35]. 08/03/2022, P. Peterlongo.
- PhD in progress: V. Epain, genome assembly with long reads, 01/10/2020, R. Andonov,D. Lavenier, JF Gibrat.
- PhD in progress: G. Gourdel, Sketch-based approaches to processing massive string data, 01/09/2020, P. Peterlongo, T. Starikovskaya.
- PhD in progress: L. Robidou, Search engine for genomic sequencing data, 01/10/2020, P. Peterlongo.
- PhD in progress: S. Romain, Genome graph data structures for Structural Variation analyses in butterfly genomes, 01/09/2021, C. Lemaitre, F. Legeai.
- PhD in progress: K. Hannoush, Pan-genome graph update strategies, 01/09/2021, P. Peterlongo, C. Marchet.

- PhD in Progress: R. Faure, Recovering end-to-end phased genomes, 01/10/2021, D. Lavenier, J-F. Flot.
- PhD in progress: M. Mognol, Processing-in-Memory, 01/04.2022, D. Lavenier.

11.2.4 Juries

- *President of PhD thesis jury.*
  - B. Churcheward, Univ. Nantes, Dec 2022 [D. Lavenier]
- *Member of PhD thesis jury*:
  - Y. Shibuya, Univ. Gustave Eiffel, Nov 2022 [K. Brinda]
  - C. Delahaye, Univ. Rennes, Dec 2022 [J. Nicolas, D. Lavenier]
- *Member of PhD thesis committee*:
  - Rick Wertenbroek, Univ. Lausanne [D. Lavenier]
  - Xavier Pic, Univ. Nice [D. Lavenier]
  - Léo de La Fuente, Univ. Rennes [D. Lavenier]
  - Khodor Hannoush, Univ. Rennes [K. Brinda]

11.3 Popularization

Participants: Victor Epain, Roland Faure, Khodor Hannoush, Garance Gourdel, Dominique Lavenier, Claire Lemaitre, Pierre Perterlongo, Sandra Romain, Lucas Robidou.

- Member of the Interstice editorial board [P. Peterlongo]
- Organization of Sciences en Cour[t]s events, Nicomaque association (link) [V. Epain, G. Gourdel, L. Robidou]

11.3.1 Articles and contents

- Short Movie "Patatogène", presented at Sciences en Courts, a local contest of popularization short movies made by PhD students (youtube video) [R. Faure, K. Hannoush, S. Romain]
- Popularization article in Interstices "Comment la bioinformatique a résolu le puzzle du génome du SARS-CoV-2" (link) [47] [C. Lemaitre]
- Popularization article in Interstices "Décoder le génome : vers la compréhension du fonctionnement du SARS-CoV-2" [48] (link) [C. Lemaitre]
- Book chapter on genome assembly, in "From Sequences to Graphs: Discrete Methods and Structures for Bioinformatics", 2022 [32] [D. Lavenier]

11.3.2 Education

- Chiche! 4 interventions in high school classes to make high school students aware of research careers in the digital sector [P. Peterlongo]

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] 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.

[8] 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.

[9] 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

[10] Y. Aigu, S. Daval, K. Gazengel, N. Marnet, C. Lariagon, A. Laperche, F. Legeai, M. J. Manzanares-Dauleux and A. Gravot. 'Multi-omic investigation of low-nitrogen conditional resistance to clubroot reveals Brassica napus genes involved in nitrate assimilation'. In: *Frontiers in Plant Science* 13 (2022), p. 790563. DOI: 10.3389/fpls.2022.790563. URL: https://hal.archives-ouvertes.fr/hal-03531145.

[11] Y. Dufresne, T. Lemane, P. Marijon, P. Peterlongo, A. Rahman, M. Kokot, P. Medvedev, S. Deorowicz and R. Chikhi. 'The K-mer File Format: a standardized and compact disk representation of sets of k-mers'. In: *Bioinformatics* 38.18 (15th Sept. 2022), pp. 4423–4425. DOI: 10.1093/bioinformatics/btac528. URL: https://hal.inria.fr/hal-03885245.

[12] R. Faure and D. Lavenier. 'QuickDeconvolution: fast and scalable deconvolution of linked-read sequencing data'. In: *Bioinformatics Advances* (27th Sept. 2022), pp. 1–8. DOI: 10.1093/bioadv/vbac068/6717790. URL: https://hal.archives-ouvertes.fr/hal-03790140.

[13] E. Fiteni, K. Durand, S. Gimenez, R. L. Meagher Jr., F. Legeai, G. J. Kergoat, N. Nègre, E. dAlençon and K. W. Nam. 'Host-plant adaptation as a driver of incipient speciation in the fall armyworm (Spodoptera frugiperda)'. In: *BMC Ecology and Evolution* 22 (11th Nov. 2022), p. 133. DOI: 10.1101/2022.09.30.510290. URL: https://hal.inrae.fr/hal-03813221.

[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, C. Lemaitre and M. Elias. 'First chromosome scale genomes of ithomiine butterflies (Nymphalidae: Ithomiini): comparative models for mimicry genetic studies'. In: *Molecular Ecology Resources* (2023). DOI: 10.1111/1755-0998.13749. URL: https://hal.inria.fr/hal-03926527.

[15] T. Lemane, R. Chikhi and P. Peterlongo. 'kmdiff, large-scale and user-friendly differential k-mer analyses'. In: *Bioinformatics* (31st Oct. 2022), pp. 1–3. DOI: 10.1093/bioinformatics/btac689. URL: https://hal.inria.fr/hal-03885124.

[16] 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://hal.inria.fr/hal-03166007.

[17] C. Meslin, P. Mainet, N. Montagné, S. Robin, F. Legeai, A. Bretaudeau, J. S. Johnston, F. A. Koutroumpa, E. Persyn, C. Monsempès, M.-C. François and E. Jacquin-Joly. 'Spodoptera littoralis genome mining brings insights on the dynamic of expansion of gustatory receptors in polyphagous noctuidae'. In: *G3* 12.8 (2nd June 2022), jkac131. DOI: 10.1093/g3journal/jkac131. URL: https://hal.inrae.fr/hal-03713321.

[18] F. Meyer, A. Fritz, Z.-L. Deng, D. Koslicki, T. R. Lesker, A. Gurevich, G. Robertson, M. Alser, D. Antipov, F. Beghini et al. 'Critical Assessment of Metagenome Interpretation: the second round of challenges'. In: *Nature Methods* 19.4 (Apr. 2022), pp. 429–440. DOI: 10.1038/s41592-022-01431-4. URL: https://hal.science/hal-03832903.

[19] D. Richter, R. Watteaux, T. Vannier, J. Leconte, P. Frémont, G. Reygondeau, N. Maillet, N. Henry, G. Benoit, A. Fernandez-Guerra, S. Suweis, R. Narci, C. Berney, D. Eveillard, F. F. Gavory, L. Guidi, K. Labadie, E. Mahieu, J. Poulain, S. Romac, S. Roux, C. Dimier, S. Kandels, M. Picheral, S. Searson, S. Pesant, J.-M. Aury, J. Brum, C. Lemaitre, E. Pelletier, P. Bork, S. Sunagawa, L. Karp-Boss, C. Bowler, M. Sullivan, E. Karsenti, M. Mariadassou, I. Probert, P. Peterlongo, P. Wincker, C. de Vargas, M. Ribera d'Alcalà, D. Iudicone and O. Jaillon. 'Genomic evidence for global ocean plankton biogeography shaped by large-scale current systems'. In: *eLife* (3rd Aug. 2022). DOI: 10.7554/eLife.78129. URL: https://hal.inria.fr/hal-02399723.

[20] E. Roux, A. Nicolas, F. Valence, G. Siekaniec, V. Chuat, J. Nicolas, Y. Le Loir and E. Guédon. 'The genomic basis of the Streptococcus thermophilus health-promoting properties'. In: *BMC Genomics* 23 (16th Mar. 2022), pp. 1–23. DOI: 10.1186/s12864-022-08459-y. URL: https://hal.inrae.fr/hal-03612308.

[21] S. Yainna, W. T. Tay, K. Durand, E. Fiteni, F. Hilliou, F. Legeai, A.-L. Clamens, S. Gimenez, R. Asokan, C. Kalleshwaraswamy, S. Deshmukh, R. Meagher, C. Blanco, P. Silvie, T. Brévault, A. Dassou, G. Kergoat, T. Walsh, K. Gordon, N. Nègre, E. d'Alençon and K. Nam. 'The evolutionary process of invasion in the fall armyworm (Spodoptera frugiperda)'. In: *Scientific Reports* 12.1 (Dec. 2022), p. 21063. DOI: 10.1038/s41598-022-25529-z. URL: https://hal.science/hal-03926301.

International peer-reviewed conferences

[22] V. Epain, R. Andonov and D. Lavenier. 'Optimal Scaffolding for Chloroplasts' Inverted Repeats'. In: JOBIM2022. Rennes, France, 5th July 2022. URL: https://hal.inria.fr/hal-03625229.

[23] G. Gourdel, A. Driemel, P. Peterlongo and T. Starikovskaya. 'Pattern matching under DTW distance'. In: *String Processing and Information Retrieval - 29th International Symposium, SPIRE 2022, Concepcion, Chile, November 8-10, 2022, Proceedings. Lecture Notes in Computer Science Springer*. SPIRE 2022 - 29th International Symposium on String Processing and Information Retrieval. Lecture Notes in Computer Science 13617. Concepción, Chile, 8th Nov. 2022, pp. 315–330. URL: https://hal.archives-ouvertes.fr/hal-03763091.

Conferences without proceedings

[24] C. Berton, G. Coatrieux and D. Lavenier. 'A first proposal for secure data storage into DNA molecules compliant with biological constraints'. In: DSMM 2022 - 1st International Conference on Data Storage in Molecular Media. Virtual, France, 21st Mar. 2022. URL: https://hal.archives-ouvertes.fr/hal-03817385.

[25] O. Boullé and D. Lavenier. 'Experimental DNA storage platform'. In: DSMM 2022 - 1st International Conference on Data Storage in Molecular Media. Virtual, France, 21st Mar. 2022, pp. 1–1. URL: https://hal.archives-ouvertes.fr/hal-03817374.

[26] V. Epain and R. Andonov. 'Linear integer programming approach for chloroplast genome scaffolding'. In: 23ème congrès annuel de la Société Française de Recherche Opérationnelle et d'Aide à la Décision. Lyon, France: 1-2, 23rd Feb. 2022. URL: https://hal.archives-ouvertes.fr/hal-03558978.

[27] R. Faure, J.-F. Flot and D. Lavenier. 'HairSplitter: assembling long reads in an unknown number of haplotypes'. In: SeqBIM 2022 - Journées sur les Séquences en Bioinformatique, Informatique et Mathématiques. Bordeaux, France, 17th Nov. 2022. URL: https://hal.archives-ouvertes.fr/hal-03864075.

[28] 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.archives-ouvertes.fr/hal-03902786.

[29] D. Lavenier. 'Processing-in-Memory to speed up NGS analysis'. In: SFA²F 2022 - Sequencing to Function: Analysis and Application for the Future. Santa Fe, United States, 21st June 2022. URL: https://hal.archives-ouvertes.fr/hal-03817360.

Scientific book chapters

[30] B. Dudek, P. Gawrychowski, G. Gourdel and T. Starikovskaya. 'Streaming Regular Expression Membership and Pattern Matching'. In: *Proceedings of the 2022 Annual ACM-SIAM Symposium on Discrete Algorithms (SODA)*. Society for Industrial and Applied Mathematics, 5th Jan. 2022, pp. 670–694. DOI: 10.1137/1.9781611977073.30. URL: https://hal.archives-ouvertes.fr/hal-03887523.

[31] C. Guyomar and C. Lemaitre. 'Metagenomics and Metatranscriptomics'. In: *From Sequences to Graphs: Discrete Methods and Structures for Bioinformatics*. ISTE, Oct. 2022. URL: https://hal.inria.fr/hal-03844316.

[32] D. Lavenier. 'Genome Assembly'. In: *From Sequences to Graphs: Discrete Methods and Structures for Bioinformatics - SCIENCES - Bioinformatics - ISTE Wiley*. 3rd Oct. 2022. URL: https://hal.archives-ouvertes.fr/hal-03817447.

Doctoral dissertations and habilitation theses

[33] C. Delahaye. 'Haplotype phasing from long reads with ASP: a flexible optimization approach'. Université rennes 1, 15th Dec. 2022. URL: https://hal.inria.fr/tel-03929660.

[34] T. Lemane. 'Indexing and analysis of large sequencing collections using kmer matrices'. Université de Rennes 1 (UR1), 16th Dec. 2022. URL: https://hal.inria.fr/tel-03921247.

[35] K. da Silva. 'Identification and quantification of microbial strains in metagenomic samples using variation graphs'. Université de Rennes 1, 8th Mar. 2022. URL: https://hal.archives-ouvertes.fr/tel-03896860.

Reports & preprints

[36] V. Epain. *Overlap Graph for Assembling and Scaffolding Algorithms: Paradigm Review and Implementation Proposals*. 2022. URL: https://hal.inria.fr/hal-03815190.

[37] V. Epain and R. Andonov. *Integer Programming Approach for Nested Pairs Genome Scaffolding*. 18th Mar. 2022. URL: https://hal.inria.fr/hal-03613353.

[38] V. Epain and R. Andonov. *Overlap Graphs for Assembling and Scaffolding Algorithms: Paradigm Review and Implementation Proposals*. 14th Oct. 2022. URL: https://hal.inria.fr/hal-03878293.

[39] V. Epain, D. Lavenier and R. Andonov. *Inverted Repeats Scaffolding for a Dedicated Chloroplast Genome Assembler*. 3rd June 2022. DOI: 10.4230/LIPIcs. URL: https://hal.inria.fr/hal-03684406.

[40] A. Guichard, F. Legeai, D. Tagu and C. Lemaitre. *MTG-Link: leveraging barcode information from linked-reads to assemble specific loci*. 28th Sept. 2022. DOI: 10.1101/2022.09.27.509642. URL: https://hal.archives-ouvertes.fr/hal-03886951.

[41] L. Robidou and P. Peterlongo. *fimpera: drastic improvement of Approximate Membership Query data-structures with counts*. 26th Dec. 2022. DOI: 10.1101/2022.06.27.497694. URL: https://hal.inria.fr/hal-03912993.

Other scientific publications

[42] C. Berton, G. Coatrieux and D. Lavenier. 'Secure data storage into DNA molecules compliant with biological constraints: Ensuring the confidentiality of data stored into DNA molecules'. In: RITS 2022 - Recherche en Imagerie et Technologies pour la Santé. Brest, France, 22nd May 2022, pp. 1–1. URL: https://hal.archives-ouvertes.fr/hal-03817968.

[43] G. Coatrieux, C. Berton, E. Dupraz, B. Hamoum, T. Derrien, Y. Audic, D. Lavenier, J. Nicolas, J. Leblanc, O. Boullé and E. Roux. 'Storing the declaration of human rights on one data DNA molecule'. In: CominLabs day 2022. Rennes, France, 10th Oct. 2022, pp. 1–1. URL: https://hal.archives-ouvertes.fr/hal-03817852.

[44] R. Faure, J.-F. Flot and D. Lavenier. 'Hairsplitter: Separating noisy long reads into an unknown number of haplotypes'. In: Genome Informatics 2022. London / Virtual, United Kingdom, 21st Sept. 2022, pp. 1–1. URL: https://hal.archives-ouvertes.fr/hal-03817928.

[45] S. Romain and C. Lemaitre. 'SVJedi-graph: genotyping close and overlapping structural variants with a variation graph and long-reads'. In: JOBIM 2022 - Journées Ouvertes en Biologie, Informatique et Mathématiques. Rennes, France, 5th July 2022. URL: https://hal.inria.fr/hal-03885541.

[46] E. Roux, A. Nicolas, F. Valence, G. Siekaniec, V. Chuat, J. Nicolas, Y. Le Loir and E. Guédon. 'The genomic basis of the Streptococcus thermophilus health-promoting properties'. In: JOBIM 2022 - Les journées Ouvertes en Biologie, Informatique et Mathématiques. Rennes, France, 8th July 2022, pp. 1–1. URL: https://hal.inrae.fr/hal-03717986.

12.3 Other

Scientific popularization

[47] C. Lemaitre, M. Salson and H. Touzet. 'Comment la bioinformatique a résolu le puzzle du génome du SARS-CoV-2'. In: *Interstices* (26th Apr. 2022). URL: https://hal.inria.fr/hal-03896532.

[48] H. Touzet, M. Salson and C. Lemaitre. 'Décoder le génome : vers la compréhension du fonctionnement du SARS-CoV-2'. In: *Interstices* (26th Apr. 2022). URL: https://hal.inria.fr/hal-03750389.

12.4 Cited publications

[49] P. Bradley, H. C. Den Bakker, E. P. Rocha, G. McVean and Z. Iqbal. 'Ultrafast search of all deposited bacterial and viral genomic data'. In: *Nature biotechnology* 37.2 (2019), pp. 152–159.

[50] P. Pandey, F. Almodaresi, M. A. Bender, M. Ferdman, R. Johnson and R. Patro. 'Mantis: a fast, small, and exact large-scale sequence-search index'. In: *Cell systems* 7.2 (2018), pp. 201–207.

[51] C. Sun, R. S. Harris, R. Chikhi and P. Medvedev. 'Allsome sequence bloom trees'. In: *Journal of Computational Biology* 25.5 (2018), pp. 467–479.