RESEARCH CENTRE

Rennes - Bretagne Atlantique

IN PARTNERSHIP WITH:

CNRS, Université Rennes 1, École normale supérieure de Rennes

2021

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

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

A3.1.2. – Data management, quering and storage

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

A3.3.3. – Big data analysis

A7.1. – 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, Researcher, HDR]
- Dominique Lavenier [CNRS, Senior Researcher, HDR]
- Claire Lemaitre [Inria, Researcher]
- Jacques Nicolas [Inria, Senior Researcher, HDR]

Faculty Members

- Roumen Andonov [Univ de Rennes I, Professor, HDR]
- Emeline Roux [Univ de Lorraine, Associate Professor, until Aug 2021]

Post-Doctoral Fellow

- Pierre Morisse [Inria, until Sep 2021]

PhD Students

- Kevin Da Silva [Inria]
- Clara Delahaye [Univ de Rennes I]
- Victor Epain [Inria/Inrae]
- Roland Faure [Univ de Rennes I, from Oct 2021]
- Garance Gourdel [Univ de Rennes I]
- Khodor Hannoush [Inria, from Sep 2021]
- Teo Lemane [Inria]
- Lucas Robidou [Inria]
- Sandra Romain [Inria, from Sep 2021]
- Gregoire Siekaniec [Institut national de recherche pour l'agriculture, l'alimentation et l'environnement, until Nov 2021]

Technical Staff

- Olivier Boulle [Inria, Engineer]
- Charles Deltel [Inria, Engineer]
- Anne Guichard [Institut national de recherche pour l'agriculture, l'alimentation et l'environnement, Engineer]
- Julien Leblanc [CNRS, Engineer, from Jul 2021]

Interns and Apprentices

- Roland Faure [Univ de Rennes I, from Mar 2021 until Aug 2021]
- Pauline Hamon-Giraud [Inria, from Apr 2021 until Jun 2021]
- Igor Martayan [CNRS, from May 2021 until Jul 2021]
- Rania Ouazahrou [Institut national de recherche pour l'agriculture, l'alimentation et l'environnement, until Jul 2021]
- Gregoire Prunier [Inria, until Jul 2021]
- Sandra Romain [Inria, until Jul 2021]
- Jordan Tayac-Geoffroy [Inria, from May 2021 until Jul 2021]

Administrative Assistant

- Marie Le Roic [Inria]

External Collaborators

- Susete Alves Carvalho [Institut national de recherche pour l'agriculture, l'alimentation et l'environnement]
- Fabrice Legeai [Institut national de recherche pour l'agriculture, l'alimentation et l'environnement]
- Emeline Roux [Univ de Rennes I, from Oct 2021]

2 Overall objectives

2.1 Genomic data processing

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), that provide up to billions of very short DNA fragments of high quality (short reads), and third generation sequencing (TGS), that provide millions of long DNA fragments of lower quality (long reads). Synthetic long reads or linked-reads is another technology type that combine the high quality and low cost of short-reads sequencing with a 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 sciences. As a matter of fact, the last sequencing machines generate Tera bytes of DNA sequences from which time-consuming processes must be applied to extract useful and pertinent information.

Today, a large number of biological questions can be investigated using genomic data. DNA is extracted from one or several living organisms, sequenced with high throughput sequencing machines, then analyzed with bioinformatics pipelines. Such pipelines are generally made of several steps. The first step performs basic operations such as quality control and data cleaning. The next steps operate more complicated tasks such as genome assembly, variant discovery (SNP, structural variations), automatic annotation, sequence comparison, etc. The final steps, based on more comprehensive data extracted from the previous ones, go toward interpretation, generally by adding different semantic information, or by performing high-level processing on these pre-processed data.

GenScale expertise relies mostly on the first and second steps. The challenge 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 that Tera bytes of raw data absolutely need to be represented 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 as short as

possible or, at least, must last a reasonable amount of time. In that case, conventional algorithms that were working on rather small datasets must be revisited to scale on today sequencing data. Parallelism is a complementary technique for increasing scalability.

GenScale research is then organized along three main axes:

– Axis 1: Data structures
– Axis 2: Algorithms
– Axis 3: Parallelism

The first axis aims at developing advanced data structures dedicated to sequencing data. Based on these objects, the second axis provides low memory footprint algorithms for a large panel of usual tools dedicated to sequencing data. Fast execution time is improved by the third axis. The combination of these three components allows efficient and scalable algorithms to be designed.

2.2 Life science partnerships

A second important objective of GenScale is to create and maintain permanent partnerships with other life science research groups. As a matter of fact, the collaboration with genomic research teams is of crucial importance for validating our tools, and for capturing new trends in the bioinformatics domain. Our approach is to actively participate in solving biological problems (with our partners) and to get involved in a few challenging genomic projects.

Partnerships are mainly supported by collaborative projects (such as ANR projects or ITN European projects) in which we act as bioinformatics partners either for bringing our expertise in that domain or for developing *ad hoc* tools.

3 Research program

3.1 Axis 1: Data Structures

The aim of this axis is to develop efficient data structures for representing the mass of genomic data generated by the sequencing machines. This research is motivated by the fact that the treatments of large genomes, such as mammalian or plant genomes, or multiple genomes coming from a same sample as in metagenomics, require high computing resources, and more specifically very important memory configuration. The advances in TGS 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, have 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 [3, 4].

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 a very 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 [1].

- **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 [3] and on the scaffolding step [5]. 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 [2].

- **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 [36].

- **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.

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 [4]. The second is a parallel algorithmic approach: during the development step, the algorithms are constrained by parallel criteria [2]. This is particularly true for parallel algorithms targeting hardware accelerators.

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.

Neurodegenerative disorders: The biological processes that lead from abnormal protein accumulation to neuronal loss and cognitive dysfunction is not fully understood. In this context, neuroimaging biomarkers and statistical methods to study large datasets play a pivotal role to better understand the pathophysiology of neurodegenerative disorders. The discovery of new genetic biomarkers could thus have a major impact on clinical trials by allowing inclusion of patients at a very early stage, at which treatments are the most likely to be effective. Correlations with genetic variables can determine subgroups of patients with common anatomical and genetic characteristics.

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

We present in this highlight an important published result regarding the error profile of Nanopore third generation sequencing technology, *Troubles and bias in Nanopore sequencing technology* [13].

This work concerns Nanopore long read sequencing, a technology that is a growing source of genomic data, since it offers a low access cost and the possibility of sequencing in the field. The counterpart is that it produces high error rate sequences compared to a short reads mature technology such as Illumina's, or even the last generation of PacBio long reads. Many articles currently focus on how to reduce this error rate after sequencing. On the other hand, the precise landscape of errors has been the subject of very little work and the technology provider, Oxford Nanopore, communicates little about the precise characteristics of its devices and softwares that are not open-source.

This paper is of interest to a wide audience of nanopore technology users. In particular, designers of software performing basecalling or assembly can take advantage of a better knowledge of the sequencer's weaknesses for their improvement. Similarly, the findings are useful for improving the analysis of variants in genomic sequences, an area where it is necessary to differentiate accurately between variations and errors. Finally, biologists can better filter their data according to quality by controlling the associated risk of error.

The technology depends on an essential software component, the basecaller, which transforms the observed electrical signal into nucleotide sequences. We propose analysis results for two generations of basecallers, including the most recent one, which show constants in the type of errors produced. The most important one concerns biases in relation to the GC rate of sequences, a characteristic not described so far but which has a proven impact on these errors. The study of a more obvious defects in homopolymer sequencing has been extended to other motifs of low complexity. Finally, we show an interesting correlation between the quality of the reads and the error rate. Our results also contain an analysis of errors for RNA direct sequencing, one of the advanced possibility of nanopores. Overall, we provide a very detailed panel of sequencing errors and this analysis can be adapted to the evolution of the technology and the data of each user thanks to a downloadable software. From an experimental point of view, this study concerns the bacterial and human genomes, and cover different contexts: prokaryotes vs. eukaryotes, genome size, genome GC levels, types of repeats. Moreover, we provide an analysis of errors for direct RNA sequencing on the Brassica napus genome.

7 New software and platforms

7.1 New software

7.1.1 MTG-link

Keywords: Bioinformatics, Genome assembly, High throughput sequencing

Functional Description: MTG-Link is a gap-filling tool for draft genome assemblies, dedicated to linked-read data generated for instance by 10X Genomics Chromium technology. It is a Python pipeline combining the local assembly tool MindTheGap and an efficient read subsampling scheme based on the barcode information of each read. It takes as input a set of reads, a GFA file with gap coordinates and an alignment file in BAM format. It outputs the results in a GFA file.

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

Publication: hal-03441914

Contact: Claire Lemaitre

Participants: Anne Guichard, Fabrice Legeai, Claire Lemaitre

Partner: INRAE

7.1.2 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 modules and a library which allows to interact with the module outputs.

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

Contact: Pierre Peterlongo

Participants: Teo Lemane, Rayan Chikhi, Pierre Peterlongo

7.1.3 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

7.1.4 StrainFLAIR

Name: STRAIN-level proFiLing using vArIation gRaph

Keywords: Indexation, Bacterial strains, Pangenomics

Functional Description: StrainFLAIR (STRAIN-level proFiLing using vArIation gRaph) is a tool for strain identification and quantification that uses a variation graph representation of gene sequences. The input is a collection of complete genomes, draft genomes or metagenome-assembled genomes from which genes will be predicted. StrainFLAIR is sub-divided into two main parts: first, an indexing step that stores clusters of reference genes into variation graphs, and then, a query step using mapping of metagenomic reads to infer strain-level abundances in the queried sample.

URL: https://github.com/kevsilva/StrainFLAIR

Contact: Kevin Da Silva

7.1.5 LRez

Keywords: High throughput sequencing, Genome analysis, Indexation

Functional Description: LRez is a C++ API and toolkit for analyzing and managing Linked-Reads sequencing data. Linked-Reads technologies, such as 10x Genomics, Haplotagging, stLFR and TELL-Seq, partition and tag high-molecular-weight DNA molecules with a barcode prior to classical short-read sequencing. This way, Linked-Reads manage to combine the high-quality of the short reads and a long-range information which can be inferred by identifying distant reads belonging to the same DNA molecule with the help of the barcodes. LRez provides various functionalities such as extracting, indexing and querying Linked-Reads barcodes, in BAM, FASTQ, and gzipped FASTQ files. The API is compiled as a shared library, helping its integration to external projects.

URL: https://github.com/morispi/LRez

Publications: hal-03421103, hal-03441917

Contact: Claire Lemaitre

Participants: Pierre Morisse, Fabrice Legeai, Claire Lemaitre

7.1.6 LEVIATHAN

Keywords: High throughput sequencing, Structural Variation, Genome analysis

Functional Description: LEVIATHAN is a structural variant calling tool dedicated to Linked-Reads sequencing data. Linked-Reads technologies combine the high quality and low cost of short-reads sequencing with a long-range information by adding barcodes that tag reads originating from the same long DNA fragment. The method relies on a barcode index, that allows to quickly compare the similarity of all possible pairs of regions in terms of amount of common barcodes. Region pairs sharing a sufficient number of barcodes are then considered as potential structural variants, and complementary, classical short reads methods are applied to further refine the breakpoint coordinates.

URL: https://github.com/morispi/LEVIATHAN

Publication: hal-03441874

Contact: Claire Lemaitre

Participants: Pierre Morisse, Fabrice Legeai, Claire Lemaitre

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.

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

Contact: Roland Faure

Partner: Université libre de Bruxelles

7.1.8 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.

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

Contact: Roland Faure

7.1.9 findere

Keywords: Indexation, Data structures, K-mer, Bloom filter, Genomic sequence

Functional Description: findere is a simple strategy for speeding up queries and for reducing false positive calls from any Approximate Membership Query data structure (AMQ). With no drawbacks (in particular no false positive), queries are two times faster with two orders of magnitude less false positive calls.

Publication: hal-03243791

Contact: Lucas Robidou

7.1.10 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

Partners: IMT Atlantique, Université de Rennes 1

7.1.11 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

8 New results

8.1 Algorithms for genome assembly and variant detection

8.1.1 Structural Variant detection with linked-reads

Participants: Fabrice Legeai, Claire Lemaitre, Pierre Morisse.

Thanks to their long-range information, linked-reads are particularly useful for structural variant calling. As a result, multiple structural variant calling methods were developed within the last few years. However, these methods were mainly tested on human data, and do not run well on non-human organisms, they all require large amounts of computing resources. We present LEVIATHAN, a new structural variant calling tool that aims to address these issues, and especially better scale and apply to a wide variety of organisms. Our method relies on a barcode index, that allows to quickly compare the similarity of all possible pairs of regions in terms of amount of common barcodes. Region pairs sharing a sufficient number of barcodes are then considered as potential structural variants, and complementary, classical short reads methods are applied to further refine the breakpoint coordinates. Our experiments on simulated data underline that our method compares well to the state-of-the-art, both in terms of recall and precision, and also in terms of resource consumption. Moreover, LEVIATHAN was successfully applied to a real dataset from a non-model organism, while all other tools either failed to run or required unreasonable amounts of resources [31].

8.1.2 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, the first SV genotyper 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. Only the SV sequences and that of the SV flanking regions are represented in our graph, resulting in a variation graph composed of multiple connected components, each representing the possible alleles for a region of one, or several close SVs. Tests on simulated long-reads on the human chromosome 1, with 1,000 deletions from the dbVar database, show a similar precision compared to SVJedi (98.1 %, against 97.8 %). Importantly, when additional deletions are added progressively closer to the original 1,000 in the dataset, SVJedi-graph maintains a 100 % genotyping rate with a high precision, when SVJedi is not able to assign a genotype to 21 % of the deletions when they are too close to each other (0-50 bp apart). SVJedi-graph also supports other SV types such as insertions and inversions, for which similar performances were obtained [40].

8.1.3 Genome gap-filling with linked-read data

Participants: Anne Guichard, Fabrice Legeai, Claire Lemaitre.

We developed a novel software, called MTG-link, for filling assembly gaps with linked-read data. This type of sequencing data has a great potential for filling the gaps as they provide long-range information while maintaining the power and accuracy of short-read sequencing. Our approach is based on local assembly using our tool MindTheGap [8], and takes advantage of barcode information to reduce the input

read set in order to reduce the de Bruijn graph complexity. MTG-Link tests different parameters values for gap-filling, followed by an automatic qualitative evaluation of the assembly. Validation was performed on a set of simulated gaps from real datasets with various genome complexities. It showed that the read subsampling step of MTG-Link enables to get better genome assemblies than using MindTheGap alone. We applied MTG-Link on 12 individual genomes of a mimetic butterfly (*H. numata*), in the Supergene ANR project context. It significantly improved the contiguity of a 1.3 Mb locus of biological interest [30].

8.1.4 Linked-read deconvolution

Participants: Roland Faure, Dominique Lavenier.

Introduced recently, linked reads technologies, such as the 10X chromium system, use microfluidics to tag multiple short reads coming from the same long (50-200 kbp) fragment with a small sequence, called barcode. Such data are cheap and easy to prepare, combining the accuracy of short-read sequencing and long-range information from the barcodes. The fact that reads with the same barcode come from the same fragment of the genome is extremely rich in information and can be used in a myriad of software. However, the same barcode may be used several times for several different fragments, complicating the analyses. We have developed QuickDeconvolution (QD) a new software for deconvoluting a set of reads sharing a barcode, i.e. separating reads coming from the different fragments. This software takes as input only the sequencing data, without the need for a reference genome. We show that QuickDeconvolution outperforms existing software in terms of accuracy, speed and scalability, making it capable of deconvolving datasets inaccessible before. In particular, we demonstrate here the first example in the literature of a successfully deconvolved animal sequencing dataset, a Drosophila melanogaster dataset of 33 Gbp [38].

8.1.5 Unzipping assembly graphs with long reads and Hi-C

Participants: Roland Faure.

Long reads and Hi-C have revolutionized the field of genome assembly as they have made highly contiguous assemblies accessible even for challenging genomes. As haploid chromosome-level assemblies are now commonly achieved for all types of organisms, phasing assemblies has become the new frontier for genome reconstruction. Several tools have already been released using long reads and/or Hi-C to phase assemblies, but they all start from a set of linear sequences and are ill-suited for non-model organisms with high levels of heterozygosity. We designed GraphUnzip, a fast, memory-efficient and flexible tool to phase assembly graphs into their constituent haplotypes using long reads and/or Hi-C data. As GraphUnzip only connects sequences that already had a potential link in the assembly graph, it yields high-quality gap-less supercontigs. To demonstrate the efficiency of GraphUnzip, we tested it on the human HG00733 and the potato *Solanum tuberosum*. In both cases, GraphUnzip yielded phased assemblies with improved contiguity [29].

8.1.6 CONSENT, long read correction and assembly polishing

Participants: Pierre Morisse.

Third-generation sequencing technologies allow to sequence long reads of tens of kbp, but with high error rates, currently capped around 10%. Self-correction is thus regularly used in long reads analysis projects. We introduce CONSENT, a new self-correction method that relies both on multiple sequence alignment and local de Bruijn graphs. To ensure scalability, multiple sequence alignment computation benefits from a new and efficient segmentation strategy, allowing a massive speedup. CONSENT compares well to the state-of-the-art, and performs better on real Oxford Nanopore data. Specifically, CONSENT

is the only method that efficiently scales to ultra-long reads. Moreover, our experiments show that error correction with CONSENT improves the quality of genome assemblies. Additionally, CONSENT implements a polishing feature, allowing to correct raw assemblies. Our experiments show that CONSENT is 2-38x times faster than other polishing tools [17].

8.1.7 Efficient reads' overlaps data structure

Participants: Victor Epain, Rumen Andonov, Dominqiue Lavenier.

One of the most frequent operations in the field of genome assembly, meta-genome assembly and/or scaffolding is sequenced reads comparison. The main purpose of this operation is to get reads' overlaps — suffix-prefix alignments, required to solve various bioinformatic issues. Because of the way the genomic sequences are read — the two complementary strands are read in reverse orientation — two reads can belong to different strands. However, this information is unknown.

The below feature (called *reverse symmetry*), is an important reads' characteristic. Denote by $\mathcal{O}$ the set of overlaps and let $(u,v) \in \mathcal{O}$ be an overlap between read u suffix and read v prefix. It is known that $(\overline{v},\overline{u}) \in \mathcal{O}$, where $\overline{u}(\overline{v})$ denotes the reverse of $u(v)$ respectively. The reverse symmetry property is commonly used to not double the reads in database, and so to represent the set of overlaps in a bi-directed graph. However, this representation (largely used by the community in the domain) increases the number of iterations over overlaps (which are oriented couples between oriented reads), and slows down the corresponding algorithms. Taking advantage of the reverse symmetry we develop in this project a novel data structure that allows to efficiently store the reads' overlaps. Iterations become faster, and it is not necessary to duplicate data any more. This new graph view permits to adapt the breadth-first search algorithm to identify inverted repeats in the sequenced genomes. This work has been presented at seqBIM2021 workshop.

8.1.8 Chloroplast scaffolding based on inverted repeat regions recovered with integer linear programming

Participants: Victor Epain, Rumen Andonov, Dominqiue Lavenier.

Chloroplasts are plastids in plants' cells known for photosynthesis metabolism. Their genomes are circular and usually form a quadripartite structure such that two unique genomic regions are separated by two inverted repeats. From a pre-assembled genome obtained witha De-Bruijn graph approach, we propose an integer linear programming strategy to extract the two inverted regions. Contigs are output with an estimated multiplicity, which corresponds to an upper bound of the number of occurrences of the contig or its reverse-complement in the solution. A contig and its reverse occurrence form an inverted pair. We model inverted repeats extraction as finding a circular path from and to a given contig, maximising the number of contiguous nested inverted pairs. We propose an integer linear programming formulation to solve this problem. Our preliminary results are very encouraging and we presented them at BiATA2021 conference [24].

In collaboration with Sven Schrinner and Gunnar Klau — respectively PhD student and Prof. at Heinrich Heine Universität, Düsseldorf, we are currently working on the NP-hardness complexity proof of this problem.

8.2 Indexing data structures and compression

8.2.1 LRez, a C++ API and toolkit for analyzing and managing Linked-Reads data

Participants: Fabrice Legeai, Claire Lemaitre, Pierre Morisse.

Linked-Reads technologies combine both the high quality and low cost of short-reads sequencing and long-range information, through the use of barcodes tagging reads which originate from a common long DNA mol- ecule. This technology has been employed in a broad range of applications including genome assembly, phasing and scaffolding, as well as structural variant calling. However, to date, no tool or API dedicated to the manipulation of Linked-Reads data exist. We introduce LRez, a C++ API and toolkit that allows easy management of Linked-Reads data. LRez includes various functionalities, for computing numbers of common barcodes between genomic regions, extracting barcodes from BAM files, as well as indexing and querying BAM, FASTQ and gzipped FASTQ files to quickly fetch all reads or alignments containing a given barcode. LRez is compatible with a wide range of Linked-Reads sequencing technologies, and can thus be used in any tool or pipeline requiring barcode processing or indexing, in order to improve their performances. In this paper, we tested the index and query subcommands on multiple files from various organisms and sequencing technologies. We notably show that the running time to query all reads with a given barcode is improved by several orders of magnitude when using LRez indexing and query implementation compared to naive index-free approaches [16, 39].

8.2.2 Large-scale kmer indexing

Participants: Téo Lemane, Pierre Peterlongo.

When indexing large collections of sequencing data, a common operation that has now been implemented in several tools (Sequence Bloom Trees and variants, BIGSI, ..) is to construct a collection of Bloom filters, one per sample. Each Bloom filter is used to represent a set of kmers which approximates the desired set of all the non-erroneous kmers present in the sample. However, this approximation is imperfect, especially in the case of metagenomics data. Erroneous but abundant kmers are wrongly included, and non-erroneous but rare ones are wrongly discarded. We propose kmtricks, a novel approach for generating Bloom filters from terabase-sized collections of sequencing data.

Our main contributions, published in [36], are 1/ an efficient method for jointly counting kmers across multiple samples, including a streamlined Bloom filter construction by directly counting hashes instead of kmers; 2/ a novel technique that takes advantage of joint counting to preserve rare kmers present in several samples, improving the recovery of non-erroneous kmers.

With our approach, we were able to index the Tara Ocean bacterial metagenomic dataset which is a difficult dataset, both in terms of size and diversity with 266 billions of distinct kmers. Such an index enables to query these raw and unassembled data with sequences of arbitrary size and thus allowing new biological analyses. In addition, our experimental results highlight that the usual yet crude filtering of rare kmers is inappropriate for this type of complex dataset.

8.2.3 Pangenome graphs for strain-level profiling of metagenomic samples

Participants: Kevin Da Silva, Pierre Peterlongo.

Current studies are shifting from the use of a single flat linear reference to a representation of multiple genomes a pangenome graph in order to exploit sequencing data from metagenomic samples.

In this context, our main contributions are 1/ a full pipeline for predicting genes from bacterial strains and for indexing them in a “variation graph”; 2/ a full pipeline for mapping unknown metagenomic reads on a so-created graph and for characterizing and evaluating the abundances of strains existing in the queried sample; 3/ a proof of concept that variation graphs may be used as a replacement of flat sequences for indexing closely related species or strains, and characterizing a sample at the strain level. These methods are implemented in the software StrainFLAIR [12].

8.2.4 A novel compressed full-text index

Participants: Garance Gourdel.

Compressed full-text indexes are very efficient but still struggle to handle some DNA readsets. In [26] we show how to use one or more assembled or partially assembled genomes as the basis for a compressed full-text index of its readset. Specifically, we build a labelled tree by taking the assembled genome as a trunk and grafting onto it the reads that align to it, at the starting positions of their alignments. Next, we compute the eXtended Burrows-Wheeler Transform (XBWT) of the resulting labelled tree and build a compressed full-text index on that. Although this index can occasionally return false positives, it is usually much more compact than the alternatives.

8.2.5 Minimizing the size of kmer indexes for Approximate Membership Queries

Participants: Lucas Robidou, Pierre Peterlongo.

In [28], we propose a simple yet efficient strategy called findere, along with its implementation, to reduce the false positive rate of any approximate membership query data structures (AMQ). The implementation of findere relies on a Bloom filter. Indeed, AMQ are widely used for representing large sets of k-mers, however they suffer from non-avoidable false-positive calls that bias methods relying on such data structures. The reduction of false positive calls by our strategy is done at query time, without any modification on the original AMQ nor generating false-negative calls and with no memory overhead. Our approach speeds up queries by a factor two. Since AMQ are usually a trade-off between space and false positive rate, findere can also be used to lower the amount of space taken by an AMQ, without increasing the false positive rate.

8.2.6 Sensible hashing techniques

Participants: Pierre Peterlongo.

In [37, 25], we extended ideas from data compression by deduplication to the Bioinformatic field. The specific problems on which we have shown our approach to be useful are the clustering of a large set of DNA strings and the search for approximate matches of long substrings, both based on the design of what we call an approximate hashing function. The outcome of the new procedure is very similar to the clustering and search results obtained by accurate tools, but in much less time and with less required memory.

8.3 Experiments with the MinION Nanopore sequencer

8.3.1 Identification of bacterial strains

Participants: Téo Lemane, Jacques Nicolas, Rania Ouazahrou, Emeline Roux, Grégoire Siekaniec.

Our aim is to provide rapid algorithms for the identification of bacteria at the finest taxonomic level. We have developed an expertise in the use of the MinION long read technology and have produced and assembled many genomes for for the lactic acid bacteria *Streptococcus thermophilus* [22] in cooperation with INRAE STLO, which have been made publicly available on the NCBI and on the Microscope platform at Genoscope.

We propose a new method of bacterial strain identification based on the assumption that a nanopore read is long enough to distinguish one strain (or group of strains) from others. This method uses a

particularly compact indexing technique of a known genome database based on a tree structure of Bloom filters. It also relies on the use of spaced seeds in order to search for sequences in the index while being less sensitive to long read substitution errors. Identification is treated as an optimization problem on a strain X kmers presence matrix and solved exactly with an ASP solver. The method is implemented in a software called ORI (Oxford nanopore Reads Identification).It has shown robust bacterial identification results on real data of *Streptococcus thermophilus* [41, 20].

ORI was further used to identify reference genomes in a complex whole piglet intestinal metagenome and to best represent meta-metagenomes. This program was initiated as a new collaboration with NuMeCan, an INRAe-INSERM-University of Rennes1 team. More than 20 bacterial species were selected (representing an abundance of more than 0.5% of the metagenome) and were described by 34 genomes selected with ORI. This work is still in progress.

8.3.2 Haplotype phasing of long reads for polyploid species

Participants: Clara Delahaye, Jacques Nicolas.

We are working on assigning the reads of a sample to their native haplotype for organisms of known polyploidy, sequenced with long read technology (Oxford Nanopore's MinION). As a first step to separate true variants from sequencing errors, we have studied the profile of sequencing errors for bacterial and human datasets. We showed that GC content is a decisive factor linked to sequencing errors. In particular, low-GC reads have almost 2% fewer errors than high-GC reads. Our work highlighted that for repeated regions (homopolymers or regions with short repeats), being the source of about half of all sequencing errors, the error profile also depends on the GC content and shows mainly deletions, although there are some reads with long insertions. Another interesting finding is that the quality measure offers valuable information on the error rate as well as the abundance of reads. [13]. We are now working on haplotype phasing of reads for di- and polyploid species. We address this problem as an optimization problem, and use Answer Set Programming to solve it. Our work focuses on reasoning on the set of possible solutions, and integrating user preferences, possibly leading to several alternative answers.

8.4 Storage on DNA

8.4.1 Error correcting code targeting nanopore sequencing

Participants: Dominique Lavenier.

We proposed a novel statistical model for DNA storage, which takes into account the memory within DNA storage error events, and follows the way nanopore sequencing works. Compared to existing channel models, the proposed model represents more accurate experimental datasets. We also proposed a full error-correction scheme for DNA storage, based on a consensus algorithm [35] and non-binary LDPC codes. Especially, we introduce a novel synchronization method which allows to eliminate remaining deletion errors after the consensus, before applying a belief-propagation LDPC decoding algorithm to correct substitution errors. This method exploits the LDPC code structure to correct deletions, and does not require adding any extra redundancy [27].

8.4.2 dnarXiv platform

Participants: Olivier Boulle, Dominique Lavenier.

We have developed an experimental platform to test or emulate the full process of writing and reading data on DNA molecules. It is composed of the following main modules : encoding, synthesis, molecule

design, sequencing, DNA data processing, decoding. It is based on a flexible software architecture where real or in-silico experimentation can be performed to test and evaluate different DNA archiving strategies.

8.4.3 Molecule design

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

One of the original features of the dnarXiv project is the use of the 3rd sequencing generation developed by Oxford Nanopore Technologies. Its main characteristic is the ability to sequence long DNA molecules. To take advantage of this technology, long DNA molecules must be used as storage support. But current synthesis technologies provide only small oligo-nucleotides (max 300nt). Thus, we are currently developing a method to assemble small synthetic DNA fragments into long molecules. The proof of concept was obtained by successfully assembling 20 single-stranded synthetic DNA fragments into a 600 bp double-stranded molecule.

8.5 Bioinformatics Analysis

8.5.1 Genomics of agro-ecosystems 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 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 [21, 10, 15, 23, 19, 18, 11, 14].

9 Bilateral contracts and grants with industry

Participants: Dominique Lavenier.

- 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 PhD CIFRE contract will start in January 2022.

10 Partnerships and cooperations

10.1 International research visitors

10.1.1 Visits to international teams

Research stays abroad

Victor Epain, PhD

Visited institution: Algorithmic Bioinformatics at l'Heinrich Heine Universität (HHU)

Country: Germany

Dates: December 1, 2021 - January 30, 2022

Context of the visit: cooperation

Mobility program/type of mobility: internship

10.2 European initiatives

10.2.1 Other european programs/initiatives

ITN IGNITE

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

- Program: ITN (Initiative Training Network)
- Project acronym: IGNITE
- Project title: Comparative Genomics of Non-Model Invertebrates
- Duration: 48 months (April 2018, March 2022)
- Coordinator: Gert Woerheide
- Partners: Ludwig-Maximilians-Universität München (Germany), Centro Interdisciplinar de Investigação Marinha e Ambiental (Portugal), European Molecular Biology Laboratory (Germany), Université Libre de Bruxelles (Belgium), University of Bergen (Norway), National University of Ireland Galway (Ireland), University of Bristol (United Kingdom), Heidelberg Institute for Theoretical Studies (Germany), Staatliche Naturwissenschaftliche Sammlungen Bayerns (Germany), INRA Rennes (France), University College London (UK), University of Zagreb (Croatia), Era7 Bioinformatics (Spain), Pensoft Publishers (Bulgaria), Queensland Museum (Australia), INRIA, GenScale (France), Institut Pasteur (France), Leibniz Supercomputing Centre of the Bayerische Akademie der Wissenschaften (Germany), Alphabiotoxine (Belgium)
- Abstract: Invertebrates, i.e., animals without a backbone, represent 95 per cent of animal diversity on earth but are a surprisingly underexplored reservoir of genetic resources. The content and architecture of their genomes remain poorly characterised, 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. IGNITE will considerably enhance our knowledge and understanding of animal genome knowledge by generating and analyzing novel data from undersampled invertebrate lineages and by developing innovative new tools for high-quality genome assembly and analysis.

ITN ALPACA

Participants: Khodor Hannoush, Pierre Peterlongo.

- Program: ITN (Innovative Training Network)
- Project acronym: ALPACA

- Project title: Comparative Genomics of Non-Model Invertebrates
- Duration: 48 months (2021-2025)
- Coordinator: Alexander Schönhuth
- Partners: Universität Bielefeld (Germany), CNRS (France), Universitaà di Pisa (Italy), Universitaà degli studi di Milano-Bicocca (Italy), Stichting Nederlandse Wetenschappelijk Onderzoek Instituten (Netherlands), Heinrich-Heine-Universität Düsseldorf (Germany), EMBL (United Kingdom), Univerzita Komenskeho v Bratislave (Slovakia), Helsingin Yliopisto (Finland), Institut Pasteur (France), The Chancellor Masters and Scholars of the University of Cambridge (United Kingdom), Geneton, s.r.o (Slovakia), Illumina Cambridge LTD, BaseClear BV, Cornell University, Whole Biome (US), Deinove (France), Suomen Punainen Risti.
- Abstract: 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. Graphbased pan-genome representations that allow to remove redundancies without having to discard individual differences, make utmost sense. In this project, we will put this shift of paradigms—from sequence to graph based representations of genomes—into full effect. As a result, we can expect a wealth of practically relevant advantages, among which arrangement, analysis, compression, integration and exploitation of genome data are the most fundamental points. In addition, we will also open up a significant source of inspiration for computer science itself. For realizing our goals, our network will (i) decisively strengthen and form new ties in the emerging community of computational pan-genomics, (ii) perform research on all relevant frontiers, aiming at significant computational advances at the level of important breakthroughs, and (iii) boost relevant knowledge exchange between academia and industry. Last but not least, in doing so, we will train a new, "paradigm-shift-aware" generation of computational genomics researchers.

10.3 National initiatives

10.3.1 ANR

Project Supergene: The consequences of supergene evolution

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

- Coordinator: M. Joron (Centre d'Ecologie Fonctionnelle et Evolutive (CEFE) UMR CNRS 5175, Montpellier)
- Duration: 48 months (Nov. 2018 – Oct. 2022)
- 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

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.

10.3.2 Inria Exploratory Action

DNA-based data storage system

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

- Coordinator : D. 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 patform allowing in-silico and real experimentations. It is a complementary work with the dnarXiv project.

10.4 Regional initiatives

10.4.1 Labex Cominlabs

dnarXiv: archiving information on DNA molecules

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

- Coordinator : D. Lavenier
- Duration : 39 months (Oct. 2020, Dec. 2023)
- Description: The dnarXiv project aims to explore 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

11 Dissemination

11.1 Promoting scientific activities

11.1.1 Scientific events: organisation

General chair

- seqBIM2021: national meeting of the sequence algorithms GT seqBIM, Lyon, Nov 2021 (2 days) [C. Lemaitre]
- (JC)2BIM: Spring school of Bioinfomatics of the GDR BIM, Rennes, Dec 2021 (5 days) [C. Lemaitre]
- JOBIM 2022: French symposium of Bioinformatics [F. Legeai]

11.1.2 Scientific events: selection

Chair of conference program committees

- JOBIM 2022: French symposium of Bioinformatics [C. Lemaitre]
- seqBIM2021: national meeting of the sequence algorithms GT seqBIM [C. Lemaitre]

Member of the conference program committees

- JOBIM 2021: French symposium of Bioinformatics [C. Lemaitre]
- CPM 2021 [P. Peterlongo]
- BIBM 2021 [D. Lavenier]
- ISMB-ECCB 2021 [D. Lavenier]

Reviewer

- ICALP 2021 [G. Gourdel]
- IWOCA 2021 [G. Gourdel]
- ISAAC 2021 [G. Gourdel]
- CPM 2021 [P. Peterlongo]
- Recomb 2021 [P. Peterlongo]
- iABC 2021 [P. Peterlongo]

11.1.3 Journal

Member of the editorial boards

- Insects [F. Legeai]

Reviewer - reviewing activities

- Nucleic Acids Research [C. Lemaitre]
- Nature Reviews Genetics [C. Lemaitre]
- Bioinformatics [P. Peterlongo, D. Lavenier]
- Journal of Experimental Algorithmics (JEA) [P. Peterlongo]
- PLOS Computational Biology [D. Lavenier]
- Molecular Ecology Resources (MER) [F. Legeai]
- Insect Biochemistry and Molecular Biology (IBMB) [F. Legeai]
- Journal of Proteomics [E. Roux]

11.1.4 Invited talks

- D. Lavenier, "Stockage d'information sur ADN", Institut Brestois du Numérique et des Mathématique, Nov. 2021
- C. Lemaitre, "Local assembly approaches for variant calling and genome assembly", Seminar of DGMI UMR, Montpellier, July 2021.

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) [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

- Member of the CoNRS, section 06, until Aug. 2021 [D. Lavenier]
- Member of the CoNRS, section 51, until Aug. 2021[D. Lavenier]
- 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]
- Representative of the environmental axis of the IRISA UMR [C. Lemaitre]
- Chair of the committee in charge of all the temporary recruitments (“Commission Personnel”) at Inria Rennes-Bretagne Atlantique and IRISA [D. Lavenier]
- Member of the Selection Committee for Lecturer Position "Maitre de Conférence" at Laboratoire IBISC (University Evry, section 27 (Informatique) [R. Andonov]

11.2 Teaching - Supervision - Juries

11.2.1 Teaching

- Licence : R. Andonov, V. Epain, Models and Algorithms in Graphs, 100h, L3, Univ. Rennes 1, France.
- Licence : G. Gourdel, Python, 48h, L2 MIASH, Univ. Paris 1, France.
- Licence : E. Roux, biochemistry, 50h, L1 and L3, Univ. Rennes 1, France.
- Master : R. Andonov, V. Epain, Operations Research (OR), 82h, M1 Miage, Univ. Rennes 1, France.
- Master : R. Andonov, Optimisation Techniques in Bioinformatics, 18h, M2, Univ. Rennes 1, France.
- Master : V. Epain, C. Lemaitre, P. Peterlongo, Algorithms on Sequences, 52h, M2, Univ. Rennes 1, France.
- Master : C. Lemaitre, T. Lemane, Bioinformatics of Sequences, 40h, M1, Univ. Rennes 1, France.
- Master : P. Peterlongo, Experimental Bioinformactics, 24h, M1, ENS Rennes, France.
- Master : F. Legeai, RNA-Seq, Metagenomics and Variant discovery, 10h, M2, National Superior School Of Agronomy, Rennes, France.
- Master : D. Lavenier, Memory Efficient Algorithms for Big Data, 24h, Engineering School, ESIR, Rennes.
- Master : D. Lavenier, Colloquium, 15h, research master degree in computer science, Univ Rennes 1
- Master : E. Roux, biochemistry, 50h, M1 and M2, Univ. Rennes 1, France.
- Aggreg: D. Lavenier, Computer Architecture, 10h, ENS Rennes
- Ecole Jeunes Chercheurs : C. Lemaitre, Genome assembly, 5h, Ecole JC2BIM du GDR BIM, Rennes

11.2.2 Defenses

- HDR: C. Lemaitre, Bioinformatics methods for studying Structural Variations with sequencing data, Université de Rennes 1, 02/12/2021 [32].
- PhD: G. Siekaniec, Identification of strains of a bacterial species from long reads, Université de Rennes 1, 10/12/2021 [33].

11.2.3 Supervision

- PhD: G. Siekaniec, Identification of strains of a bacterial species from long reads, J. Nicolas (co-supervised with E. Guédon, E. Roux).
- PhD in progress: K. da Silva, Metacatalogue : a new framework for intestinal microbiota sequencing data mining, 01/10/2018, P. Peterlongo (co-supervised with M. Berland, N. Pons).
- PhD in progress: C. Delahaye, Robust interactive reconstruction of polyploid haplotypes, 01/10/2019, J. Nicolas.
- PhD in progress: T. Lemane, unbiased detection of neurodegenerative structural variants using k-mer matrices, 01/10/2019, P. Peterlongo.
- PhD in progress: V. Epain, Genome Assembly with Long Reads, 01/10/2020 R. Andonov, D. Lavenier, (co-supervised with JF Gibrat, INRAE).
- PhD in progress: G. Gourdel, Sketch-based approaches to processing massive string data, 01/09/2020, P. Peterlongo (co-supervised with 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, D. Lavenier, C. Lemaitre.
- PhD in progress: K. Hannoush, Pan-genome graph update strategies, 01/09/2021, P. Peterlongo (co-supervised with C. Marchet).
- PhD in Progress: R. Faure, Recovering end-to-end phased genomes, 01/10/2021, D. Lavenier (co-supervised with J-F. Flot).

11.2.4 Juries

- *Member of Habilitation thesis jury*: C. Lemaitre [D. Lavenier, president]
- *Referee of Ph-D thesis jury*: Vincent Sater, Univ Rouen [P. Peterlongo], Y. Mansour, Univ. Montpellier [D. Lavenier]
- *Member of PhD thesis jury*: Quentin Delorme, Univ Montpellier [C. Lemaitre], Camille Sessegolo, Univ Lyon [P. Peterlongo], Chi Nguyen Lam, UBO [D. Lavenier].
- *Member of PhD thesis committee*: Benoit Goutorbe, Univ Paris-Saclay [C. Lemaitre] Benjamin Churcheward, Univ. Nantes [D. Lavenier], Belaid Hamoum, UBS, Lorient [D. Lavenier], Nguyen Dang, Univ. Montpellier [D. Lavenier], Rick Wertenbroek, Univ. Lausanne [D. Lavenier], Xavier Pic, Univ. Nice [D. Lavenier].

11.3 Popularization

11.3.1 Internal or external Inria responsibilities

- Member of the Interstice editorial board [P. Peterlongo]
- Organization of Sciences en cour[t]s events, Nicomaque association (link) [C. Delahaye, T. Lemane]

11.3.2 Articles and contents

- Short Movie "Cocktails de bio-informatique", presented at Sciences en Courts, a local contest of popularization short movies made by PhD students (link) [G. Gourdel, V. Epain, L. Robidou]
- Popularization report from the GDR BIM, "SARS-CoV-2 Through the Lens of Computational Biology: How bioinformatics is playing a key role in the study of the virus and its origins" [34] [C. Lemaitre]

12 Scientific production

12.1 Major publications

[1] 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.

[2] 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.

[3] 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.

[4] 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.

[5] S. François, R. Andonov, D. Lavenier and H. Djidjev. 'Global optimization approach for circular and chloroplast genome assembly'. In: *BICoB 2018 - 10th International Conference on Bioinformatics and Computational Biology*. Las Vegas, United States, Mar. 2018, pp. 1–11. DOI: 10.1101/231324. URL: https://hal.inria.fr/hal-01666830.

[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] G. Bianchetti, V. Clouet, F. Legeai, C. Baron, K. Gazengel, A. Carrillo, M. M. Manzanares-Dauleux, J. J. Buitink and N. Nesi. 'RNA sequencing data for responses to drought stress and/or clubroot infection in developing seeds of Brassica napus'. In: *Data in Brief* 38 (Oct. 2021), pp. 1–11. DOI: 10.1016/j.dib.2021.107392. URL: https://hal-univ-rennes1.archives-ouvertes.fr/hal-03379739.

[11] A. Cusumano, S. Urbach, F. Legeai, M. Ravallec, M. Dicke, E. Poelman and A.-N. Volkoff. 'Plant-phenotypic changes induced by parasitoid ichnoviruses enhance the performance of both unparasitized and parasitized caterpillars'. In: *Molecular Ecology* 30.18 (Sept. 2021), pp. 4567–4583. DOI: 10.1111/mec.16072. URL: https://hal.archives-ouvertes.fr/hal-03287280.

[12] K. Da Silva, N. Pons, M. Berland, F. Plaza Oñate, M. Almeida and P. Peterlongo. 'StrainFLAIR: strain-level profiling of metagenomic samples using variation graphs'. In: *PeerJ* (23rd Aug. 2021). DOI: 10.7717/peerj.11884. URL: https://hal.inria.fr/hal-03141144.

[13] C. Delahaye and J. Nicolas. 'Sequencing DNA with nanopores: Troubles and biases'. In: *PLoS ONE* (1st Oct. 2021), pp. 1–29. DOI: 10.1371/journal.pone.0257521. URL: https://hal.inria.fr/hal-03362956.

[14] J.-L. Gatti, M. Belghazi, F. Legeai, M. Ravallec, M. FRAYSSINET, S. Robin, D. Aboubakar-Souna, R. Srinivasan, M. Tamò, M. Poirié and A.-N. Volkoff. 'Proteo-Trancriptomic Analyses Reveal a Large Expansion of Metalloprotease-Like Proteins in Atypical Venom Vesicles of the Wasp Meteorus pulchricornis (Braconidae)'. In: *Toxins* 13.7 (19th July 2021), pp. 1–36. DOI: 10.3390/toxins13070502. URL: https://hal.inrae.fr/hal-03292170.

[15] J. Gauthier, H. Boulain, J. J. F. A. van Vugt, L. Baudry, E. Persyn, J.-M. Aury, B. Noel, A. Bretaudeau, F. Legeai, S. Warris et al. 'Chromosomal scale assembly of parasitic wasp genome reveals symbiotic virus colonization'. In: *Communications Biology* 4.1 (26th Jan. 2021), pp. 1–15. DOI: 10.1038/s42003-020-01623-8. URL: https://hal.archives-ouvertes.fr/hal-03127732.

[16] P. Morisse, C. Lemaitre and F. Legeai. 'LRez: C++ API and toolkit for analyzing and managing Linked-Reads data'. In: *Bioinformatics Advances* 1.1 (9th June 2021), pp. 1–4. DOI: 10.1093/bioadv/vbab022. URL: https://hal.inria.fr/hal-03421103.

[17] P. Morisse, C. Marchet, A. Limasset, T. Lecroq and A. Lefebvre. 'Scalable long read self-correction and assembly polishing with multiple sequence alignment'. In: *Scientific Reports* 11.1 (Dec. 2021), pp. 1–13. DOI: 10.1038/s41598-020-80757-5. URL: https://hal-cnrs.archives-ouvertes.fr/hal-03210290.

[18] F. Piron-Prunier, E. Persyn, F. Legeai, M. McClure, C. Meslin, S. Robin, S. Alves-carvalho, A. Mohammad, C. Blugeon, E. Jacquin-joly, N. Montagné, M. Elias and J. Gauthier. 'Comparative transcriptome analysis at the onset of speciation in a mimetic butterfly—The Ithomiini Melinaea marsaeus'. In: *Journal of Evolutionary Biology* 34.11 (Nov. 2021), pp. 1704–1721. DOI: 10.1111/jeb.13940. URL: https://hal.archives-ouvertes.fr/hal-03381525.

[19] E. Poivet, A. Gallot, N. Montagné, P. Senin, C. Monsempès, F. Legeai and E. Jacquin-Joly. 'Transcriptome Profiling of Starvation in the Peripheral Chemosensory Organs of the Crop Pest Spodoptera littoralis Caterpillars'. In: *Insects* 12.7 (23rd June 2021), p. 573. DOI: 10.3390/insects12070573. URL: https://hal.sorbonne-universite.fr/hal-03278298.

[20] G. Siekaniec, E. Roux, T. Lemane, E. Guédon and J. Nicolas. 'Identification of isolated or mixed strains from long reads: a challenge met on Streptococcus thermophilus using a MinION sequencer'. In: *Microbial Genomics* 7.11 (2021), pp. 1–14. DOI: 10.1099/mgen.0.000654. URL: https://hal.archives-ouvertes.fr/hal-03444296.

[21] K. S. Singh, E. Cordeiro, B. Troczka, A. Pym, J. Mackisack, T. Mathers, A. Duarte, F. Legeai, S. Robin, P. Bielza, H. Burrack, K. Charaabi, I. Denholm, C. Figueroa, R. ffrench-Constant, G. Jander, J. Margaritopoulos, E. Mazzoni, R. Nauen, C. Ramírez, G. Ren, I. Stepanyan, P. Umina, N. Voronova, J. Vontas, M. Williamson, A. Wilson, G. Xi-Wu, Y.-N. Youn, C. Zimmer, J.-C. Simon, A. Hayward and C. Bass. 'Global patterns in genomic diversity underpinning the evolution of insecticide resistance in the aphid crop pest Myzus persicae'. In: *Communications Biology* 4.1 (Dec. 2021), p. 847. DOI: 10.1038/s42003-021-02373-x. URL: https://hal.inrae.fr/hal-03313531.

[22] O. Uriot, M. Kebouchi, E. Lorson, W. Galia, S. Denis, S. Chalancon, Z. Hafeez, E. Roux, M. Genay, S. BLANQUET-DIOT and A. Dary-Mourot. 'Identification of Streptococcus thermophilus Genes Specifically Expressed under Simulated Human Digestive Conditions Using R-IVET Technology'. In: *Microorganisms* 9.6 (21st May 2021), pp. 1–26. DOI: 10.3390/microorganisms9061113. URL: https://hal.inrae.fr/hal-03269961.

International peer-reviewed conferences

[23] S. Alves Carvalho, K. Gazengel, A. Bretaudeau, S. Robin, S. Daval and F. Legeai. 'AskoR, A R Package for Easy RNASeq Data Analysis'. In: IECE 2021 - 1st International Electronic Conference on Entomology. Virtual, France, 2021, pp. 1–8. DOI: 10.3390/IECE-10646. URL: https://hal.inrae.fr/hal-03347665.

[24] R. Andonov, V. Epain and D. Lavenier. 'Optimal de novo assemblies for chloroplast genomes based on inverted repeats patterns'. In: Bioinformatics: from Algorithms to Applications 2021. St. Petersbourg, Russia, France, 12th July 2021. URL: https://hal.inria.fr/hal-03534195.

[25] G. Arbitman, S. T. Klein, P. Peterlongo and D. Shapira. 'Approximate Hashing for Bioinformatics'. In: *LNCS*. CIAA 2021 - 25th International Conference on Implementation and Application of Automata. Vol. 12803. 25th International Conference on Implementation and Application of Automata. Bremen, Germany, 19th July 2021, pp. 1–12. URL: https://hal.inria.fr/hal-03219482.

[26] T. Gagie, G. Gourdel and G. Manzini. 'Compressing and Indexing Aligned Readsets'. In: WABI 2021 - Workshop on Algorithms in Bioinformatics. Online conference, France, 2nd Aug. 2021, pp. 1–21. DOI: 10.4230/LIPIcs.WABI.2021.13. URL: https://hal.archives-ouvertes.fr/hal-03478058.

[27] B. Hamoum, E. Dupraz, L. Conde-Canencia and D. Lavenier. 'Channel Model with Memory for DNA Data Storage with Nanopore Sequencing'. In: ISTC 2021 - 11th International Symposium on Topics in Coding. Montreal, Canada: IEEE, 30th Aug. 2021, pp. 1–5. URL: https://hal-imt-atlantique.archives-ouvertes.fr/hal-03337117.

[28] L. Robidou and P. Peterlongo. 'findere: fast and precise approximate membership query'. In: SPIRE 2021 - The 28th annual Symposium on String Processing and Information Retrieval. Lille / Virtual, France, 4th Oct. 2021. DOI: 10.1101/2021.05.31.446182. URL: https://hal.inria.fr/hal-03243791.

Conferences without proceedings

[29] R. Faure, N. Guiglielmoni and J.-F. Flot. 'GraphUnzip: unzipping assembly graphs with long reads and Hi-C'. In: JOBIM 2021 - Journées Ouvertes en Biologie, Informatique et Mathématiques. Paris, France, 6th July 2021, pp. 1–7. URL: https://hal.archives-ouvertes.fr/hal-03441016.

[30] A. Guichard, F. Legeai, D. Tagu and C. Lemaitre. 'MTG-Link: filling gaps in draft genome assemblies with linked read data'. In: JOBIM 2021 - Journées Ouvertes Biologie, Informatique et Mathématiques. Paris, France, 6th July 2021, pp. 1–8. URL: https://hal.inria.fr/hal-03441914.

[31] P. Morisse, F. Legeai and C. Lemaitre. 'LEVIATHAN: efficient discovery of large structural variants by leveraging long-range information from Linked-Reads data'. In: JOBIM 2021 - Journées Ouvertes en Biologie, Informatique et Mathématiques. Paris, France, 6th July 2021, pp. 1–8. URL: https://hal.inria.fr/hal-03441874.

Doctoral dissertations and habilitation theses

[32] C. Lemaitre. 'Méthodes bioinformatiques pour l'étude des Variants de Structure avec des données de séquençages génomiques'. Université Rennes 1, 2nd Dec. 2021. URL: https://tel.archives-ouvertes.fr/tel-03497793.

[33] G. R. Siekaniec. 'Identification of strains of a bacterial species from long reads'. MathSTIC, 10th Dec. 2021. URL: https://tel.archives-ouvertes.fr/tel-03510672.

Reports & preprints

[34] S. Alizon, F. Cazals, S. Guindon, C. Lemaitre, T. Mary-Huard, A. Niarakis, M. Salson, C. Scornavacca and H. Touzet. *SARS-CoV-2 Through the Lens of Computational Biology:How bioinformatics is playing a key role in the study of the virus and its origins*. CNRS, 15th Mar. 2021, pp. 1–35. URL: https://hal-cnrs.archives-ouvertes.fr/hal-03170023.

[35] D. Lavenier. *Constrained Consensus Sequence Algorithm for DNA Archiving*. CNRS IRISA, 11th May 2021. URL: https://hal.archives-ouvertes.fr/hal-03479083.

[36] T. Lemane, P. Medvedev, R. Chikhi and P. Peterlongo. *kmtricks: Efficient construction of Bloom filters for large sequencing data collections*. 11th Mar. 2021. DOI: 10.1101/2021.02.16.429304. URL: https://hal.inria.fr/hal-03166007.

Other scientific publications

[37] G. Arbitman, S. T. Klein, P. Peterlongo and D. Shapira. 'Approximate Hashing for Bioinformatics'. In: DCC 2021 - Data Compression Conference. Virtual, United States, 24th Mar. 2021. URL: https://hal.inria.fr/hal-03166000.

[38] R. Faure. 'QuickDeconvolution: fast and scalable deconvolution of linked-reads sequencing data'. Sorbonne universités, 6th Sept. 2021. URL: https://hal.inria.fr/hal-03479233.

[39] P. Morisse, C. Lemaitre and F. Legeai. 'LRez: C++ API and toolkit for analyzing and managing Linked-Reads data'. In: JOBIM 2021 - Journées Ouvertes en Biologie, Informatique et Mathématiques. Paris, France, 6th July 2021, p. 1. URL: https://hal.inria.fr/hal-03441917.

[40] S. Romain and C. Lemaitre. 'SVJedi-graph: Structural Variant genotyping with long-reads using a variation graph'. In: JOBIM 2021 - Journées Ouvertes en Biologie, Informatique et Mathématiques. Paris, France, 6th July 2021, p. 1. URL: https://hal.inria.fr/hal-03441915.

[41] G. Siekaniec, R. Ouazahrou, G. Boudry, E. Guédon, E. Roux and J. Nicolas. 'Identification of bacterial strains using ORI (Oxford nanopore Reads Identification)'. In: Microbes 2021 - Société Française de Microbiologie. Nantes, France, 22nd Sept. 2021. URL: https://hal.archives-ouvertes.fr/hal-03374572.