RESEARCH CENTRE

Rennes - Bretagne Atlantique

IN PARTNERSHIP WITH:

CNRS, Université Rennes 1, École normale supérieure de Rennes

2020

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 Team: 2012 January 01, updated into 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

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

- Dominique Lavenier [Team leader, CNRS, Senior Researcher, HDR]
- Pierre Peterlongo [Team leader, Inria, 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]

Post-Doctoral Fellows

- Celine Le Beguec [Inria, until Jun 2020]
- Pierre Morisse [Inria, from Sep 2020]

PhD Students

- Kevin Da Silva [Inria]
- Wesley Delage [Inria]
- Clara Delahaye [Univ de Rennes I]
- Victor Epain [Inria, from Oct 2020]
- Garance Gourdel [Univ de Rennes I, from Sep 2020]
- Lolita Lecompte [Inria, until Nov 2020]
- Teo Lemane [Inria]
- Lucas Robidou [Inria, from Oct 2020]
- Gregoire Siekaniec [Institut national de recherche pour l'agriculture, l'alimentation et l'environnement]

Technical Staff

- Olivier Boulle [Inria, Engineer, from Oct 2020]
- Charles Deltel [Inria, Engineer]
- Anne Guichard [Institut national de recherche pour l'agriculture, l'alimentation et l'environnement, Engineer]

Interns and Apprentices

- Emmanuel Clostres [Univ de Rennes I, from Apr 2020 until Jul 2020]
- Victor Epain [Inria, from Feb 2020 until Jul 2020]
- Thomas Rigole [Inria, from Feb 2020 until Sep 2020]
- Kerian Thuillier [Inria, from May 2020 until Jul 2020]

Administrative Assistant

- Marie Le Roic [Inria]

Visiting Scientist

- Zeyuan Chen [INRA, until Feb 2020]

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]
- Mohammed-Amin Madoui [CEA, until Mar 2020]

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 billions of very short DNA fragments of high quality, and third generation sequencing (TGS), that provide millions of long DNA fragments of lower quality. NGS and TGS techniques 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 scalable 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 last 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].

- **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. See Section 6.1.

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 drug. 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

6.1 Project dnarXiv

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. Our main objective in this project 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 we will consider two key promising biotechnologies: enzymatic DNA synthesis and DNA nanopore sequencing. We will 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.

The dnarXiv project is starting this year as it obtained several fundings, with an Inria AEx and a CominLabs projects. It involves several research teams (from Irisa, Lab-STICC Université Bretagne Sud, Inserm, IGDR UMR 6290) and an industrial partner (DNAScript).

7 New software and platforms

7.1 New software

7.1.1 SVJedi

Keywords: High throughput sequencing, Structural Variation, Genome analysis

Functional Description: SVJedi is a structural variation (SV) genotyper for long read data. Based on a representation of the different alleles, it estimates the genotype of each variant in a given individual sample based on allele-specific alignment counts. SVJedi 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/llecompte/SVJedi

Contacts: Claire Lemaitre, Lolita Lecompte

Participants: Claire Lemaitre, Lolita Lecompte, Pierre Peterlongo, Dominique Lavenier

7.1.2 MinYS

Name: MineYourSymbiont

Keywords: High throughput sequencing, Genome assembly, Metagenomics

Functional Description: MinYS allows targeted assembly of a bacterial genome of interest in a metagenomic short read sequencing sample using a reference-guided pipeline. First, taking advantage of a potentially distant reference genome, a subset of the metagenomic reads is assembled into a set of backbone contigs. Then, this first draft assembly is completed using the whole metagenomic readset in a de novo manner. The resulting assembly is output as a genome graph, allowing to distinguish different strains with potential structural variants coexisting in the sample.

URL: https://github.com/cguyomar/MinYS

Contacts: Claire Lemaitre, Cervin Guyomar

7.1.3 Simka

Keywords: Comparative metagenomics, K-mer, Distance, Ecology

Functional Description: Simka is a comparative metagenomics method dedicated to NGS datasets. It computes a large collection of distances classically used in ecology to compare communities by approximating species counts by k-mer counts. The method scales to a large number of datasets thanks to an efficient and parallel kmer-counting strategy that processes all datasets simultaneoulsy. SimkaMin is distributed also with Simka. SimkaMin is a faster and more resource-frugal version of Simka. It outputs approximate (but very similar) results by subsampling the kmer space.

Release Contributions: Since release version 1.5.0, SimkaMin is distributed alongside Simka. SimkaMin is also a de novo comparative metagenomics tool. It is a faster and more resource-frugal version of Simka. It outputs approximate (but very similar) results as Simka by subsampling the kmer space. With this strategy, and with default parameters, SimkaMin is an order of magnitude faster, uses 10 times less memory and 70 times less disk than Simka.

URL: https://gatb.inria.fr/software/simka/

Publications: hal-01595071, hal-02308101

Authors: Gaëtan Benoit, Claire Lemaitre, Pierre Peterlongo

Contacts: Gaëtan Benoit, Claire Lemaitre

Participants: Claire Lemaitre, Dominique Lavenier, Gaëtan Benoit, Pierre Peterlongo, Charles Deltel

7.1.4 DiscoSnpRad

Name: DISCOvering Single Nucleotide Polymorphism, Indels in RAD seq data

Keyword: RAD-seq

Functional Description: Software discoSnpRad is designed for discovering Single Nucleotide Polymorphism (SNP) and insertions/deletions (indels) from raw set(s) of RAD-seq data. Note that number of input read sets is not constrained, it can be one, two, or more. Note also that no other data as reference genome or annotations are needed. The software is composed of several modules. First module, kissnp2, detects SNPs from read sets. A second module, kissreads2, enhances the kissnp2 results by computing per read set and for each variant found i/ its mean read coverage and ii/ the (phred) quality of reads generating the polymorphism. Then, variants are grouped by RAD locus, and a VCF file is finally generated. We also provide several scripts to further filter and select informative variants for downstream population genetics studies.

This tool relies on the GATB-Core library.

Release Contributions: * Substantive improvements: better quality of results (accuracy and recall), better filtering of obtained results * Formal improvements: better organization of scripts, better presentation of results

URL: https://github.com/GATB/DiscoSnp

Contact: Pierre Peterlongo

Participants: Pierre Peterlongo, Claire Lemaitre

7.1.5 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

Contacts: Claire Lemaitre, Anne Guichard, Fabrice Legeai

Partner: INRAE

7.1.6 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

Authors: Teo Lemane, Rayan Chikhi, Pierre Peterlongo

Contacts: Teo Lemane, Pierre Peterlongo

Participants: Teo Lemane, Rayan Chikhi, Pierre Peterlongo

7.1.7 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

Authors: Gregoire Siekaniec, Teo Lemane, Jacques Nicolas

Contacts: Gregoire Siekaniec, Jacques Nicolas

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

7.1.8 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

Contacts: Kevin Da Silva, Pierre Peterlongo

8 New results

8.1 Algorithms for genome assembly and variant detection

8.1.1 Small variant discovery in RAD-like data

Participants Fabrice Legeai, Claire Lemaitre, Pierre Peterlongo.

Restriction site Associated DNA Sequencing (RAD-Seq) is a technique characterized by the sequencing of specific loci along the genome, that is widely employed in the field of evolutionary biology since it allows to exploit variants (mainly Single Nucleotide Polymorphism—SNPs) information from entire populations at a reduced cost, and without relying on a reference genome. Common RAD dedicated tools are based on all-vs-all read alignments, which require consequent time and computing resources. We present an original method, DiscoSnp-RAD, that avoids this pitfall since variants are detected by looking for specific motifs in the de Bruijn graph built from the whole read sets. We tested the implementation on simulated datasets of increasing size, up to 1,000 samples, and on real RAD-Seq data from 259 specimens of Chiastocheta flies, morphologically assigned to seven species. DiscoSnp-RAD allowed all individuals to be successfully assigned to their species. Moreover, identified variants succeeded to reveal a within-species genetic structure linked to the geographic distribution. Furthermore, our results show that DiscoSnp-RAD is significantly faster than state-of-the-art tools, in particular on large datasets [19].

In the context of the SPECREP ANR project, DiscoSnpRAD was then used on several mimetic butterfly species. It allowed the comparison of the genetic structure and hybridization patterns of two mimetic butterfly species, *Ithomia salapia* and *Oleria onega* (Nymphalidae: Ithomiini), each consisting of a pair of lineages differentiated for their wing colour pattern and that come into contact in the Andean foothills of Peru. Our results highlight major differences, both at the genomic and phenotypic level, between the two species and contrast with the genomic patterns observed for other well studied mimetic butterlflies [20].

8.1.2 Structural Variation genotyping

Participants Dominique Lavenier, Lolita Lecompte, Claire Lemaitre, Pierre Peterlongo.

Structural variations (SV) are genomic variants of at least 50 base pairs (bp) that can be rearranged within the genome and thus can have a major impact on biological processes. Sequencing data from third generation technologies have made it possible to better characterize SVs. One of the problems is the genotyping of variants. It consists in estimating the presence and ploidy or absence of a set of known variants in a newly sequenced individual. Although many SV callers have been published recently, there was no published method to date dedicated to genotyping SVs with this type of data. We present here a novel method and its implementation, SVJedi, to genotype SVs with long reads. From a set of known SVs and a reference genome, our approach first generates local sequences representing the two possible alleles for each SV. Long read data are then aligned to these generated sequences and a careful analysis of the alignments consists in identifying only the informative ones to estimate the genotype for each SV. SVJedi achieves high accuracy on simulated and real human data and we demonstrate its substancial benefits with respect to other existing approaches, namely SV discovery with long reads and SV genotyping with short reads [24].

8.1.3 Genome assembly of targeted organisms in metagenomic data

Participants Wesley Delage, Fabrice Legeai, Claire Lemaitre.

In this work, we propose a two-step targeted assembly method tailored for metagenomic data, called MinYS (for MineYourSymbiont). First, a subset of the reads belonging to the species of interest are recruited by mapping and assembled *de novo* into backbone contigs using a classical assembler. Then an all-versus-all contig gap-filling is performed using a novel version of MindTheGap [8] with the whole metagenomic dataset. The originality and success of the approach lie in this second step, that enables to assemble the missing regions between the backbone contigs, which may be regions absent or too divergent from the reference genome. The result of the method is a genome assembly graph in gfa format, accounting for the potential structural variations identified within the sample. We showed that MinYS is able to assemble the *Buchnera aphidicola* genome in a single contig in pea aphid metagenomic samples, even when using a divergent reference genome. It runs at least 10 times faster than classical *de novo* metagenomics assemblers and it is able to recover large structural variations co-existing in a sample [23].

8.1.4 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. In linked-read technologies, such as the 10X Genomics Chromium platform, reads that have been sequenced from the same long DNA molecule (30-50 Kb) can be identified by a small barcode sequence. Thus, these technologies have 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 [43, 42].

8.2 Indexing data structures

8.2.1 A probabilistic data structure for indexing large sets of kmers

Participants Lolita Lecompte, Pierre Peterlongo.

Indexing massive data sets is extremely expensive for large scale problems. This work proposes a probabilistic data structure based on a minimal perfect hash function for indexing large sets of keys. Our structure out-competes the hash table for construction and query times and for memory usage, in the case of the indexation of a static set. To illustrate the impact of these performances, we provide two applications based on similarity computation with kmers between collections of sequences, and for which this calculation is an expensive but required operation. In particular, we show a practical case in which other bioinformatics tools fail to scale up the tested data set or provide results of lower quality [26].

8.2.2 Large-scale kmer indexation

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 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. In addition, our experimental results highlight that the usual yet crude filtering of rare kmers is inappropriate for complex data such as metagenomes.

The work had been done in 2020, the paper was written late 2020 and submitted early 2021.

8.2.3 SimkaMin: subsampling the kmer space for efficient comparative metagenomics

Participants Claire Lemaitre, Pierre Peterlongo.

SimkaMin is a quick comparative metagenomics tool with low disk and memory footprints, thanks to an efficient data subsampling scheme used to estimate Bray-Curtis and Jaccard dissimilarities. One billion metagenomic reads can be analyzed in less than 3 minutes, with tiny memory (1.09 GB) and disk (~0.3 GB) requirements and without altering the quality of the downstream comparative analyses, making of SimkaMin a tool perfectly tailored for very large-scale metagenomic projects [12].

8.2.4 Indexing and querying 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 as pangenome graphs 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.

The work had been done in 2020, the paper was written late 2020 and submitted early 2021.

8.3 Theoretical results

8.3.1 Linked-read study

Participants Dominique Lavenier.

Considering a set of intervals on the real line, an interval graph records these intervals as nodes and their intersections as edges. Identifying pairs of nodes in an interval graph results in a multiple-interval graph. Given only the nodes and the edges of the multiple-interval graph without knowing the underlying intervals, we are interested in the following questions. Can one determine how many intervals correspond to each node? Can one compute a walk over the multiple-interval graph nodes that reflects the ordering of the original intervals? These questions are closely related to linked-read

DNA sequencing, where barcodes are assigned to long molecules whose intersection graph forms an interval graph. Each barcode may correspond to multiple molecules, which complicates downstream analysis, and corresponds to the identification of nodes of the corresponding interval graph. Resolving the above graph-theoretic problems would facilitate the analyses of linked-read sequencing data, through enabling the conceptual separation of barcodes into molecules and providing, through the molecule order, a skeleton for accurately assembling the genome. Here, we propose a framework that takes as input an arbitrary intersection graph and constructs a heuristic approximation of the ordering of the original intervals [30]. This research is done in collaboration with R. Chikhi, Pasteur Institute.

8.3.2 Hardness results of some pattern mining problems

Participants Garance Gourdel.

We initiate a study on the fundamental relationship between data sanitization (i.e., the process of hiding confidential information in a given dataset) and frequent pattern mining, in the context of sequential (string) data to address the fact that current methods introduce a number of spurious patterns that may harm the utility of frequent pattern mining. We present several hardness results on variants of the problem, essentially showing that these variants cannot be solved or even be approximated in polynomial time and we propose integer linear programming formulations to solve them, which work in polynomial time under certain realistic assumptions on the problem parameters. Those solutions have been implemented and tested [32].

8.4 Optimization

8.4.1 Integer Linear Programming for de novo long read assembly

Participants Rumen Andonov, Victor Epain, Dominique Lavenier.

Long read sequencing technologies offer the possibility to overcome the major issue in genome assembly, namely the genome's repeated regions. But they suffer from a high error rate, with sequencing errors, like nucleotide insertions or deletions, called indels. Although some long-read assemblers already exist according to several methods, such as using De Bruijn graphs or correcting iteratively the reads for example, we propose here a different approach based on mixed integer linear programming (MILP). This is a two steps strategy : first, we attribute to the maximum of reads a position on a same position axis and then we produce a consensus sequence thanks to the positioning. At these aims, we propose a modelling for the positioning issue with the mixed integer linear programming (MILP), and we present the first ideas for the consensus sequence production and multiple sequences alignment from the positioning, with MILP too. As the final aim of this strategy is to formalize the genome assembly problem, we structured it according to the mathematical method, that permits to target methodological choices precisely, and then reducing the heuristic uses. Finally, we tested the strategy with bacteria genomes. Despite the fact that positioning results are positive, the consensus results are mitigated but do not remove the potentiality of the method association [41, 38].

8.4.2 Genome haplotyping with short paired reads

Participants Rumen Andonov, Mohammed Amin Madoui, Pierre Peterlongo, Kerian Thuillier.

A biological sample contains often more than one individual or comes from a polyploid organism, an organism which has several versions of its genome. For example, human beings are diploids, they have two versions of the human genome: one version inherited from the father and the other inherited from

the mother. The different versions of one genome are called haplotypes. During the sequencing process, all the haplotypes are sequenced together. Reconstructing all the haplotypes from a set of reads is called *haplotype-aware genome assembly*.

We propose two new approaches for haplotype-aware genome assembly based on Mixed Integer Linear Programming for multicommodity-flow problems. Unlike previous approaches, we use a global optimization scheme and avoid the use of heuristics. The input data are generated by the *DiscoSNP* algorithm [9]. *DiscoSNP* is a *de novo* variant detection tool. Here, a variant is a single nucleotide difference between several haplotypes.

A comparison between both approaches is performed and the advantages and limits of each one are discussed. We also proposed an experimental protocol to test the quality of the solution [45].

8.5 Parallelism

8.5.1 Variant detection using a processing-in-memory technology

Participants Dominique Lavenier.

The concept of Processing-In-Memory aims to dispatch the computer power near the data. Together with the UPMEM company (http://www.upmem.com/), which is currently developing a DRAM memory enhanced with computing units, we parallelized the detection of small mutations on the human genome. Traditionally, this process is split into 2 steps: a mapping step and a variant calling step. Here, thanks to the high processing power of this new type of memory, the mapping step can nearly be done at the disk transfer rate. We have defined an ad-hoc data structure allowing the variant calling step to be performed simultaneously on the host processor. Basically, the two steps are overlapped in such a way that reads are mapped by packet. When a packet is mapped, the mapping results of the previous one dynamically feed the variant calling data structure. Our variant calling implementation on real UPMEM systems show a large speed-up compared to a standard multithreaded software. Compared with other acclerators (GPU or FPGA), the execution time is identical, but energy consumption is strongly reduced [33].

8.6 Experiments with the MinION Nanopore sequencer

8.6.1 Identification of bacterial strains

Participants Téo Lemane, Jacques Nicolas, 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 lactic bacteria in cooperation with INRAE STLO, which have been made publicly available on the NCBI (https://www.ncbi.nlm.nih.gov/) and on the Microscope platform at Genoscope (http://www.genoscope.cns.fr/agc/microscope/). We have developed a first classifier called ORI (Oxford nanopore Reads Identification) that demonstrates the possibility to identify isolated strains with spaced seed indexing of the noisy long reads produced by the MinION. [44]

8.6.2 Studying the sequencing error profile of the nanopore sequencer

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 [40].

8.6.3 Information archiving on DNA molecules

Participants Olivier Boulle, Dominique Lavenier, Jacques Nicolas, Emeline Roux.

In 2020, we started a new research project, called dnarXiv, whose aim is to investigate the archiving of information on DNA molecules. Indeed, DNA offers the advantage of a storage density of 1,000 to 10,000 times greater than current technologies. The objective is to set up a prototype to demonstrate the feasibility of this new storage possibility based on the latest bio-technologies: enzymatic synthesis (writing information) and sequencing nanopore (reading information). From a numerical point of view, the research axes focus on how to encode information inside DNA molecules and on how to ensure their safety. This project is supported by the Labex CominLab (https://project.inria.fr/dnarxiv/) and by an Inria's Eploratory Action.

8.7 Benchmarks and Reviews

8.7.1 Evaluation of error correction tools for long read data

Participants Lolita Lecompte, Pierre Peterlongo.

Long read technologies, such as Pacific Biosciences and Oxford Nanopore, have high error rates (from 9% to 30%). Hence, numerous error correction methods have been recently proposed, each based on different approaches and, thus, providing different results. As this is important to assess the correction stage for downstream analyses, we designed the ELECTOR software, providing evaluation of long read correction methods. This software generates additional quality metrics compared to previous existing tools. It also scales to very long reads and large datasets and is compatible with a wide range of state-of-the-art error correction tools [27].

8.7.2 Evaluation of insertion variant callers on real human data

Participants Wesley Delage, Claire Lemaitre.

Insertion variants are one of the most common types of structural variation. Although such variants have many biological impacts on species evolution and health, they have been understudied because they are very difficult to detect with short read re-sequencing data. Recently, with the commercialization of novel long read technologies, insertion variants are finally being discovered and referenced in human populations. Thanks to several international efforts, some gold standard call sets have been produced in 2019, referencing tens of thousands insertions. On these datasets, all existing short-read insertion variant callers, including our own method MindTheGap [8], can reach at most 5 to 10 % of the referenced insertion variants. In this work, we propose a precise characterization of the different types of insertion variants, based on the nature and size of their inserted sequence, the genomic context of their insertion site and the complexity at their breakpoints. In a detailed benchmark, we then analyze which of these features impact most the recall of existing methods. By simulating the identified factors of difficulty, we investigate the causes of low recall and how these can be bypassed or improved in existing algorithms. [14, 31]

8.7.3 Artificial Intelligence and Bioinformatics

Participants Jacques Nicolas.

We took part in the writing of an ambitious book on artificial intelligence in 3 volumes, "A Guided Tour of Artificial Intelligence Research", edited by P. Marquis, O. Papini, and H. Prade. The chapter we wrote shines a light on the strong links shared by Artificial intelligence and Bioinformatics since many years. We introduced a selection of the challenging problems for Artificial intelligence offered by Bioinformatics as well as an extensive bibliography of the state-of-the-art. Together with the framing of questions, we point to several achievements and progresses made in the literature with the hope it can help the bioinformatician, bioanalyst or biologist to have access to state of the art methods [34].

8.8 Bioinformatics Analysis

8.8.1 Genome assembly and analysis of a parasitic wasp and its integrated viral genome

Participants Fabrice Legeai, Claire Lemaitre.

We used Illumina paired-end and mate-pair reads to assemble the genome of a parasitic wasp, *Hyposoter didymator*, carrying integrated viral particules, and we annotated it with RNA-Seq data. We identified all the viral insertions in the genome and located precisely the excision sites of a set of circularized viral sequences by developing a specific method called DrjBreakpointFinder and freely distributed at `http://github.com/stephanierobin/DrjBreakpointFinder/`. Finally we studied the colinearity between *Hyposoter didymator* and the genome of another ichnovirus carrying wasp (*Campoletis sonorensis*). These analyses and comparisons of the viral insertion structures in these two wasps helped elucidating the mechanisms that have facilitated viral domestication in ichneumonid wasps. Their genomic architectures clearly differ from the organization of viral insertions in braconid wasps, revealing different evolutionary trajectories that have led to virus domestication in the two wasp lineages [25].

8.8.2 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 (`http://bipaa.genouest.org`), 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.

Concerning aphid organisms, we conducted various analyses including differential gene expression, genomics diversity from populations, and epigenomics variability among morphs of pea aphids to study the effect of selection of genes duplications all along the pea aphid genome [18]. We participated to the characterization of secreted KQY proteins, a very peculiar family of genes including a very high number of small exons sparsely located on the aphids genomes and transcribed in multiple isoforms [17]. We also participated in the annotation of the aphid *Aphis glycines*, a major pest of the soybean [22].

The fall armyworm (*Spodoptera frugiperda*) is one of the most damaging pest insects of different classes of crop plants. We produced a new release of its genome using novel data (long reads and Hi-C data) and migrated the annotation from the previous release. It allows our collaborators to annotate structural variants (CNVs, mainly duplications and deletions) among broad populations of moths and to

study the effect of the selection on the CNV harboring detoxifaction genes related to insecticide resistance [21].

We assembled the genome of *Aphidius ervi* a parasitic wasp of aphids. We used a hybrid approach mixing short reads (Illumina paired-end and mate-pair reads) and long PacBio reads. Then we identified the genes and transposable elements in the resulting scaffolds and performed the same analyses to the genome of *Lysiphlebus fabarum*, assembled by collaborators. These two genomes have provided insights into adaptive evolution in parasitoids that infect aphids [15].

We reported the genome of *Daktulosphaira vitifoliae*, the grape phylloxera. We used a similar hybrid approach mixing short and long PacBio reads for genome assembly and we annotated the genes and transposable elements. We identified an interesting expansion of a family of effector genes. Various population genomics were also performed, allowing us to test various demographic scenarios for the introduction of phylloxera in Europe. We conclude that phylloxera populations of the upper Mississippi River basin, feeding on the wild species *Vitis riparia*, are likely to be the principal source of the invasion to Europe [28].

8.8.3 Structural genome analysis of *S. pyogenes* strains

Participants Dominique Lavenier, Thomas Rigole, Emeline Roux.

The *S. pyogenes* bacteria is responsible for many human infections. With the increase in the prevalence of infections (750 million infections per year worldwide and 4th in terms of mortality from bacterial infection), a better understanding of adaptive and evolutionary mechanisms at play in this bacteria is essential. The molecular characterization of the different strains is done by the *emm* gene. A statistical analysis of the different types of *emm* on the Brittany population shows 3 main dynamics: sporadic types, endemic types or epidemic types. The last case was observed in Brittany for the type *emm75* between 2009 and 2017. Two hypotheses can be considered: (1) the emergence of a new subtype or winning clone in an unimmunized population; (2) increased pathogenicity through genetic evolution of the strains, including the acquisition of new virulence factors. In collaboration with the microbiology department of the Rennes Hospital, we sequenced more than 30 *S. pyogenes emm75* strains (Oxford Nanopore MinION sequencing) in order to study the dynamic of the epidemic thought their structural genomic variations [16].

8.8.4 Logical reasoning to study metabolic pathways

Participants Jacques Nicolas.

Inferring genome-scale metabolic networks in emerging model organisms is challenged by incomplete biochemical knowledge and partial conservation of biochemical pathways during evolution. A given phenotype can be conserved even if the underlying molecular mechanisms are changing. We formalized in a logical framework (Answer Set Programming) the inference of new reactions and molecular structures, based on previous biochemical knowledge. We have developed for this purpose a form of analogical reasoning that has allowed us to abstract and extend the known molecular transformations. This method was applied to study the metabolic pathway drift in the red algal model *Chondrus crispus* [11].

8.8.5 Detection of cytosine methylation of mature microRNAs

Participants Pierre Peterlongo.

Literature reports that mature microRNA (miRNA) can be methylated at adenosine, guanosine and cytosine. However, the molecular mechanisms involved in cytosine methylation of miRNAs have not yet

been fully elucidated. Here, we investigated the biological role and underlying mechanism of cytosine methylation in miRNAs in glioblastoma multiforme (GBM). The GenScale team provided the algorithmic solution and implementation for the detection of methylated cytosine on mature microRNAs [13].

8.8.6 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 [39].

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.

9 Bilateral contracts and grants with industry

The UPMEM company is currently developing new memory devices with embedded computing power (`http://www.upmem.com/`). GenScale investigates how bioinformatics and genomics algorithms can benefit from these new types of memory. In 2020, we parallelized the detection of short variants.

10 Partnerships and cooperations

10.1 International research visitors

10.1.1 Visits of international scientists

- Visit of Zeyuan Chen from Ludwig-Maximilians-Universitat, Munich, Germany, January 2020

10.2 European initiatives

10.2.1 Collaborations in European programs, except FP7 and H2020

ITN IGNITE

- 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

- Program: ITN (Initiative 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).
- 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. Graph-based pan-genome representations that allow to remove redundancies without having to discard individual differences, make utmost sense. In this project, we will put this shift of paradigms—from sequence to graph based representations of genomes—into full effect. As a result, we can expect a wealth of practically relevant advantages, among which arrangement, analysis, compression, integration and exploitation of genome data are the most fundamental points. In addition, we will also open up a significant source of inspiration for computer science itself. For realizing our goals, our network will (i) decisively strengthen and form new ties in the emerging community of computational pan-genomics, (ii) perform research on all relevant frontiers, aiming at significant computational advances at the level of important breakthroughs, and (iii) boost relevant knowledge exchange between academia and industry. Last but not least, in doing so, we will train a new, “paradigm-shift-aware” generation of computational genomics researchers.

10.2.2 Collaborations with major European organizations

- Two years France-Bulgaria bilateral Partnership Hubert Curien (PHC) RILA 2019 (project code : 43196Q). The topic of this project is "Integer Programming Approaches for Long-Reads Genome Assembly". Start year: 2019.

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: https://www.cesgo.org/seqdigger/

10.3.2 PIA: Programme Investissement d'Avenir

RAPSODYN: Optimization of the rapeseed oil content under low nitrogen

Participants Dominique Lavenier, Claire Lemaitre, Pierre Peterlongo.

- Coordinator: N. Nesi (INRAE, IGEPP, Rennes)
- Duration: 99 months (2012-2020)
- Partners: 5 companies, 9 academic research labs.
- Description: The objective of the Rapsodyn project is the optimization of the rapeseed oil content and yield under low nitrogen input. GenScale is involved in the bioinformatics work package to elaborate advanced tools dedicated to polymorphism detection and their application to the rapeseed plant.
- Website: http://www.rapsodyn.fr

10.3.3 Programs from research institutions

Inria Project Lab: Neuromarkers

Participants Dominique Lavenier, Céline Le Beguec, Claire Lemaitre, Téo Lemane, Pierre Peterlongo.

- Coordinator: O. Colliot (Inria, Aramis, Paris)
- Duration: 4 years (2017-2020)
- Partners: INRIA (Aramis, Pasteur, Dyliss, GenScale, XPOP), ICM
- Description: The Neuromarkers IPL aims to design imaging bio-markers of neuro-degenerative diseases for clinical trials and study of their genetic associations. In this project, GenScale brings its expertise in the genomics field. More precisely, given a case-control population, a first step is to identify small genetic variations (SNPs, small indels) from their genomes. Then, using these variations together with brain images (also partitioned into case-control data sets), the challenge is to select variants that present potential correlation with brain images.

Inria Exploratory Action: dnarXiv, archiving information on DNA molecules

Participants Olivier Boulle, Dominique Lavenier, Jacques Nicolas, Emeline Roux.

- 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. This Inria's Exploratory Action has the same objective as the CominLabs project of the same name.

10.4 Regional initiatives

Labex CominLabs: Project DNA-Store: Advanced error correction scheme for DNA-based data storage using nanopore technology

Participants Dominique Lavenier, Emeline Roux.

- Coordinator: L. Conde-Canencia (UBS, Lab-STCC, IAS)
- Duration: 12 months (Feb. 2019 - Feb. 2020)
- Partners: UBS (Lab-STCC, IAS, L. Conde-Canencia)
- Description: The DNA-Store project is funded by the Labex CominLabs. The goal is to explore the possibility to store information on DNA molecules. As DNA sequencing (the reading process) is performed with the Oxford Nanopore technology, powerful error correcting codes need to be developed together with dedicated genomic data processing.

Labex CominLabs: dnarXiv, archiving information on DNA molecules

Participants Olivier Boulle, Dominique Lavenier, Jacques Nicolas, Emeline Roux.

- Coordinator : D. Lavenier
- Duration : 39 months (Oct. 2020, dec. 2023)
- Description: This CominLabs project is the follow-up of the DNA-Store CominLabs project, and it has the same objective as the Inria's Exploratory Action. This goal 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. Advanced solutions will be proposed in terms of coding schemes (i.e., source and channel coding) and data security (i.e., data confidentiality/integrity and DNA storage authenticity), considering the constraints and advantages of the chemical processes and biotechnologies involved in DNA storage.

Project Thermin: Differential characterization of strains of a bacterial species, Streptococcus thermophilus, with a Nanopore MinION

Participants Dominique Lavenier, Jacques Nicolas, Emeline Roux, Grégoire Siekaniec.

- Coordinator: J. Nicolas (Inria/Irisa, GenScale, Rennes)
- Duration: 36 months (Oct. 2018 – Sept. 2021)
- Partners: INRAE (STLO, Agrocampus Rennes, E. Guédon and Y. Le Loir).
- Description: The Thermin project aims at exploring the capacities of a low cost third generation sequencing device, the Oxford Nanopore MinION, for rapid and robust pan-genome discrimination of bacterial strains and their phenotypes. It started with the recruitments of E. Roux (délégation INRIA, Oct, 2018), a biochemist from Lorraine University, and G. Siekaniec (INRAE -INRIA collaboration, INRAE grant), a new PhD student. We study pan-genomic representations of multiple genomes and the production of characteristic signatures of each genome in this context.

11 Dissemination

11.1 Promoting scientific activities

11.1.1 Scientific events: organisation

- The GenScale team organized and hosted the international workshop DSB 2020: "Data Structure in Bioinformatics".

11.1.2 Scientific events: selection

Chair of conference program committees

- seqBIM 2020 : national meeting of the sequence algorithms GT seqBIM [C. Lemaitre]

Member of the conference program committees

- 2020 IEEE International Conference on Bioinformatics and Biomedicine (BIBM'2020) [D. Lavenier]
- 11th International Workshop on Biological Knowledge Discovery and Data Mining (BIOKDD'20) [D. Lavenier]
- 8th International Work-Conference on Bioinformatics and Biomedical Engineering (IWBBIO 2020) [D. Lavenier]
- Symposium on Experimental Algorithms June 2020 (SEA 2020) [P. Peterlongo]

11.1.3 Journal

Member of the editorial boards

- Insects [F. Legeai]

Reviewer - reviewing activities

- Bioinformatics [D. Lavenier, P. Peterlongo]
- BMC Bioinformatics [D. Lavenier]
- BMC Medical Genomics [P. Peterlongo]
- BMC Genomics [F. Legeai]
- Genome Biology [D. Lavenier]
- Genomics [D. Lavenier]
- Frontiers in Physiolgy [F. Legeai]
- Frontiers in Microbiology [F. Legeai]
- IEEE's Transactions on Parallel and Distributed Systems [D. Lavenier]
- Journal of Computational Biology [D. Lavenier]
- Journal of Parallel and Distributed Computing [D. Lavenier]
- Molecular Ecology Resources [F. Legeai]
- PLOS Computational Biology [D. Lavenier]

11.1.4 Invited talks

- Rencontres transdiscipliaires Technologies et Santé : ADN, polymères et bigdata, oct. 2020, Paris: "Archivage d'information sur ADN" [D. Lavenier]
- Sémimaire équipe Bonsai, Cristal, Lille, mars 2020: "Post Processing of Minion Data" [D. Lavenier]
- Microbiomes Day, Oct 1st 2020, Rennes: "Identification and quantification of strains in a strain mixture using variation graphs" [K. Da Silva]
- International conference Bioinformatics: from Algorithms to Applications: Biata jul 2020, St Petersburg. Keynote Speaker: "Indexing large and numerous sequencing datasets" [P. Peterlongo]

11.1.5 Leadership within the scientific community

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

- Expert for the MEI (International Expertise Mission), French Research Ministry [D. Lavenier]
- Member of the Scientific Council of BioGenOuest [D. Lavenier]
- Member of the Scientific Council of Agrocampus Ouest (Institute for life, food and horticultural sciences and landscaping) [J. Nicolas]

11.1.7 Research administration

- Member of the CoNRS, section 06, [D. Lavenier]
- Member of the CoNRS, section 51, [D. Lavenier]
- Corresponding member of COERLE (Inria Operational Committee for the assesment of Legal and Ethical risks). Participation to the ethical group of IFB (French Elixir node, Institut Français de Bioinformatique) [J. Nicolas]
- Scientific Advisor of The GenOuest Platform (Bioinformatics Resource Center BioGenOuest) and deputy representative of Inria for the grouping council that defines Biogenouest's strategy and ensures the consistency of its actions. [J. Nicolas]
- Representative of the environnemental axis of the IRISA UMR [C. Lemaitre]
- In charge of the bachelor's degree in the computer science department of University of Rennes 1 (120 students) [R. Andonov]

11.2 Teaching - Supervision - Juries

11.2.1 Teaching

- Licence : C. Delahaye, Python, 12 h, L1, Univ. Rennes 1, France.
- Licence : R. Andonov, V. Epain, Graph Algorithms, 100h, L3, Univ. Rennes 1, France.
- License : G. Gourdel, Python, 48h, L2 MIASH, Univ. Paris 1, France.
- Master : R. Andonov, V. Epain, Operational research, 82h, M1 Miage, Univ. Rennes 1, France.
- Master : G. Siekaniec, Python, 24 h, M1, Univ. Rennes 1, France.
- Master : C. Delahaye, 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 : R. Andonov, Advanced Algorithmics, 25h, Univ. Rennes 1, France.
- Master : D. Lavenier, Memory Efficient Algorithms for Big Data, 24h, Engineering School, ESIR, Rennes.

11.2.2 Supervision

- PhD : L. Lecompte, Structural Variant genotyping with long-read sequencing data, Université de Rennes, 04/12/2020 [36].
- PhD : W. Delage, Characterization and detection of large constitutional insertions for medical use, Université de Rennes, 11/12/2020 [35].
- PhD in progress: K. da Silva, Metacatalogue : a new framework for intestinal microbiota sequencing data mining, 01/10/2018, M. Berland, N. Pons and P. Peterlongo.
- PhD in progress: G. Siekaniec, Differential characterization of strains of bacterial species, 01/10/2018, E. Guédon, E. Roux and J. Nicolas.
- 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 D. Lavenier, R. Andonov, 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

11.2.3 Juries

- *Referee of Habilitation thesis jury.*
  - D. Vallenet, Univ Evry-Val-d'Essonne [P. Peterlongo]
- *Member of PhD thesis jury.*
  - Guillaume Gautreau, Univ Paris-Saclay [C. Lemaitre],
  - Lolita Lecompte, Univ Rennes [D. Lavenier, C. Lemaitre],
  - Wesley Delage, Univ Rennes [C. Lemaitre],
  - Nicolas Bloyet, Univ Bretagne Sud [R. Andonov]
- *Member of PhD thesis committee.*
  - Benoit Goutorbe, Univ Paris-Saclay [C. Lemaitre],
  - Hugo Talibart, Univ Rennes [J. Nicolas],
  - B. Hamoum, Univ Bretagne Sud [D. Lavenier]
  - B. Churchward, Univ Nantes [D. Lavenier]
  - C. Nguyen Lam, Univ Bretagne Occidentale [D. Lavenier]
  - N. Dang, Univ Montpellier [D. Lavenier]

11.3 Popularization

11.3.1 Internal or external Inria responsibilities

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

11.3.2 Articles and contents

- Contribution to a small book of the CNRS research group in artificial intelligence (GDR IA) providing an elementary introduction to artificial intelligence, the diversity of its techniques and the variety of simulated intelligent capabilities. Our own contribution focuses on the analysis of the links between bioinformatics and AI [46].

11.3.3 Interventions

- "Chaplotype", presented at Sciences en Courts, a local contest of popularization movies made by PhD students (http://sciences-en-courts.fr/) [C. Delahaye, T. Lemane].
- L Codent L Créent : promoting programming for secondary schoolgirls during workshops supervised by female PhD students (http://lclc-rennes.irisa.fr/) [C. Delahaye].
- Fête de la science 2020, presentation of algorithmic solutions proposed for comparing seawater metagenomes. "Sciences en Direct, un ocean de richesses" [P. Peterlongo].

11.3.4 Internal actions

- Organization of Sciences en cour[t]s events, Nicomaque association (http://sciences-en-courts.fr/) [W. Delage, G.Siekaniec]

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] J. N. Alanko, H. Bannai, B. Cazaux, P. Peterlongo and J. Stoye. 'Finding all maximal perfect haplotype blocks in linear time'. In: *Algorithms for Molecular Biology* (10th Feb. 2020), pp. 1–9. DOI: 10.1186/s13015-020-0163-6. URL: https://hal.inria.fr/hal-02187246.

[11] A. Belcour, J. Girard, M. Aite, L. Delage, C. Trottier, C. Marteau, C. J.-J. Leroux, S. M. Dittami, P. Sauleau, E. Corre, J. Nicolas, C. Boyen, C. Leblanc, J. Collén, A. Siegel and G. V. Markov. 'Inferring Biochemical Reactions and Metabolite Structures to Understand Metabolic Pathway Drift'. In: *iScience* 23.2 (21st Feb. 2020), p. 100849. DOI: 10.1016/j.isci.2020.100849. URL: https://hal.inria.fr/hal-01943880.

[12] G. Benoit, M. Mariadassou, S. Robin, S. Schbath, P. Peterlongo and C. Lemaitre. 'SimkaMin: fast and resource frugal de novo comparative metagenomics'. In: *Bioinformatics* 36.4 (15th Feb. 2020), pp. 1–2. DOI: 10.1093/bioinformatics/btz685. URL: https://hal.inria.fr/hal-02308101.

[13] M. Cheray, A. Etcheverry, C. Jacques, R. Pacaud, G. Bougras-Cartron, M. Aubry, F. Denoual, P. Peterlongo, A. Nadaradjane, J. Briand, F. Akcha, D. Heymann, F. M. Vallette, J. Mosser, B. Ory and P.-F. Cartron. 'Cytosine methylation of mature microRNAs inhibits their functions and is associated with poor prognosis in glioblastoma multiforme'. In: *Molecular Cancer* 19.1 (2020), pp. 1–36. DOI: 10.1186/s12943-020-01155-z. URL: https://hal-univ-rennes1.archives-ouvertes.fr/hal-02500622.

[14] W. J. Delage, J. Thevenon and C. Lemaitre. 'Towards a better understanding of the low recall of insertion variants with short-read based variant callers'. In: *BMC Genomics* 21.762 (4th Nov. 2020). DOI: 10.1186/s12864-020-07125-5. URL: https://hal.inria.fr/hal-03032763.

[15] A. Dennis, G. Ballesteros, S. Robin, L. Schrader, J. Bast, J. Berghöfer, L. Beukeboom, M. Belghazi, A. Bretaudeau, J. Buellesbach, E. Cash, D. Colinet, Z. Dumas, M. Errbii, P. Falabella, J.-L. Gatti, E. Geuverink, J. Gibson, C. Hertaeg, S. Hartmann, E. J. Jacquin-Joly, M. Lammers, B. Lavandero, I. Lindenbaum, L. Massardier-Galatà, C. Meslin, N. Montagné, N. Pak, M. Poirié, R. Salvia, C. C. A. Smith, D. Tagu, S. Tares, H. Vogel, T. Schwander, J.-C. Simon, C. Figueroa, C. Vorburger, F. Legeai and J. Gadau. 'Functional insights from the GC-poor genomes of two aphid parasitoids, "Aphidius ervi" and "Lysiphlebus fabarum"'. In: *BMC Genomics* 21.Art. 376 (Dec. 2020), pp. 1–27. DOI: 10.1186/s12864-020-6764-0. URL: https://hal.inrae.fr/hal-02649295.

[16] M. Devaere, S. Boukthir, S. Moullec, E. Roux, D. Lavenier, A. Faili and S. Kayal. 'Complete Genome Sequences of Two Strains of Streptococcus pyogenes Belonging to an Emergent Clade of the Genotype emm89 in Brittany, France'. In: *Microbiology Resource Announcements* 9.11 (2020), e00129–20. DOI: 10.1128/MRA.00129-20. URL: https://hal-univ-rennes1.archives-ouvertes.fr/hal-02532959.

[17] M. Dommel, J. Oh, J. C. Huguet-Tapia, E. Guy, H. Boulain, A. Sugio, M. Murugan, F. Legeai, M. Heck, M. C. Smith and F. White. 'Big Genes, Small Effectors: Pea Aphid Cassette Effector Families Composed From Miniature Exons'. In: *Frontiers in Plant Science* 11 (2nd Sept. 2020). DOI: 10.3389/fpls.2020.01230. URL: https://hal.inria.fr/hal-03065308.

[18] R. Fernández, M. Marcet-Houben, F. Legeai, G. Richard, S. Robin, V. Wucher, C. Pegueroles, T. Gabaldón and D. Tagu. 'Selection following Gene Duplication Shapes Recent Genome Evolution in the Pea Aphid Acyrthosiphon pisum'. In: *Molecular Biology and Evolution* 37.9 (1st Sept. 2020), pp. 2601–2615. DOI: 10.1093/molbev/msaa110. URL: https://hal.inria.fr/hal-03065353.

[19] J. Gauthier, C. Mouden, T. Suchan, N. Alvarez, N. Arrigo, C. Riou, C. Lemaitre and P. Peterlongo. 'DiscoSnp-RAD: de novo detection of small variants for RAD-Seq population genomics'. In: *PeerJ* (10th June 2020), pp. 1–20. DOI: 10.7717/peerj.9291. URL: https://hal.inria.fr/hal-01634232.

[20] J. Gauthier, D. L. De-Silva, Z. Gompert, A. Whibley, C. Houssin, Y. Le Poul, M. McClure, C. Lemaitre, F. Legeai, J. Mallet and M. Elias. 'Contrasting genomic and phenotypic outcomes of hybridization between pairs of mimetic butterfly taxa across a suture zone'. In: *Molecular Ecology* 29.7 (Apr. 2020), pp. 1328–1343. DOI: 10.1111/mec.15403. URL: https://hal.archives-ouvertes.fr/hal-02800429.

[21] S. Gimenez, H. Abdelgaffar, G. L. Goff, F. Hilliou, C. Blanco, S. Hänniger, A. Bretaudeau, F. Legeai, N. Nègre, J. L. Jurat-Fuentes, E. D'Alençon and K. Nam. 'Adaptation by copy number variation increases insecticide resistance in the fall armyworm'. In: *Communications Biology* 3.1 (Dec. 2020), p. 664. DOI: 10.1038/s42003-020-01382-6. URL: https://hal.inria.fr/hal-03065279.

[22] R. Giordano, R. K. Donthu, A. Zimin, I. C. Julca Chavez, T. Gabaldon, M. Van Munster, L. Hon, R. Hall, J. Badger, M. Nguyen et al. 'Soybean aphid biotype 1 genome: Insights into the invasive biology and adaptive evolution of a major agricultural pest'. In: *Insect Biochemistry and Molecular Biology* 120 (May 2020), p. 103334. DOI: 10.1016/j.ibmb.2020.103334. URL: https://hal.inria.fr/hal-03065382.

[23] C. Guyomar, W. Delage, F. Legeai, C. Mougel, J.-C. Simon and C. Lemaitre. 'MinYS: mine your symbiont by targeted genome assembly in symbiotic communities'. In: *NAR Genomics and Bioinformatics* 2.3 (1st Sept. 2020), pp. 1–11. DOI: 10.1093/nargab/lqaa047. URL: https://hal.inria.fr/hal-02891885.

[24] L. Lecompte, P. Peterlongo, D. Lavenier and C. Lemaitre. 'SVJedi: genotyping structural variations with long reads'. In: *Bioinformatics* 36.17 (1st Sept. 2020), pp. 4568–4575. DOI: 10.1093/bioinformatics/btaa527. URL: https://hal.inria.fr/hal-03032737.

[25] F. F. Legeai, B. F. Santos, S. Robin, A. Bretaudeau, R. B. Dikow, C. Lemaitre, V. Jouan, M. Ravallec, J.-M. Drezen, D. Tagu, F. Baudat, G. Gyapay, X. Zhou, S. Liu, B. A. Webb, S. Brady and A.-N. Volkoff. 'Genomic architecture of endogenous ichnoviruses reveals distinct evolutionary pathways leading to virus domestication in parasitic wasps'. In: *BMC Biology* 18.1 (Dec. 2020), pp. 1–23. DOI: 10.1186/s12915-020-00822-3. URL: https://hal.inrae.fr/hal-02918230.

[26] C. Marchet, L. Lecompte, A. Limasset, L. Bittner and P. Peterlongo. 'A resource-frugal probabilistic dictionary and applications in bioinformatics'. In: *Discrete Applied Mathematics* 92-102.Volume 274 (15th Mar. 2020). DOI: 10.1016/j.dam.2018.03.035. URL: https://hal.inria.fr/hal-01322440.

[27] C. Marchet, P. Morisse, L. Lecompte, A. Lefebvre, T. Lecroq, P. Peterlongo and A. Limasset. 'ELECTOR: evaluator for long reads correction methods'. In: *NAR Genomics and Bioinformatics* 2.1 (1st Mar. 2020), pp. 1–12. DOI: 10.1093/nargab/lqz015. URL: https://hal.inria.fr/hal-02371117.

[28] C. Rispe, F. Legeai, P. Nabity, R. Fernández, A. Arora, P. Baa-Puyoulet, C. Banfill, L. Bao, M. Barberà, M. Bouallegue et al. 'The genome sequence of the grape phylloxera provides insights into the evolution, adaptation, and invasion routes of an iconic pest'. In: *BMC Biology* 18.1 (Dec. 2020), p. 90. DOI: 10.1186/s12915-020-00820-5. URL: https://hal.inrae.fr/hal-02917617.

[29] J. A. Wenger, B. J. Cassone, F. Legeai, J. S. Johnston, R. Bansal, A. D. Yates, B. S. Coates, V. A. C. Pavinato and A. Michel. 'Whole genome sequence of the soybean aphid, Aphis glycines.' In: *Insect Biochemistry and Molecular Biology* 123 (1st Aug. 2020), p. 102917. DOI: 10.1016/j.ibmb.2017.01.005. URL: https://hal.inria.fr/hal-01555244.

International peer-reviewed conferences

[30] Y. Dufresne, C. Sun, P. Marijon, D. Lavenier, C. Chauve and R. Chikhi. 'A Graph-Theoretic Barcode Ordering Model for Linked-Reads'. In: WABI 2020 - 20th Workshop on Algorithms in Bioinformatics. Pisa, Italy, 20th Aug. 2020, pp. 11–12. DOI: 10.4230/LIPIcs.WABI.2020.11. URL: https://hal.archives-ouvertes.fr/hal-03008334.

National peer-reviewed Conferences

[31] W. Delage, J. Thevenon and C. Lemaitre. 'Towards a better understanding of the low discovery rate of short-read based insertion variant callers'. In: JOBIM 2020 - Journées Ouvertes Biologie, Informatique et Mathématiques. Montpellier, France, 30th June 2020. URL: https://hal.inria.fr/hal-03120668.

Conferences without proceedings

[32] G. Bernardini, A. Conte, G. Gourdel, R. Grossi, G. Loukides, N. Pisanti, S. P. Pissis, G. Punzi, L. Stougie and M. Sweering. 'Hide and Mine in Strings: Hardness and Algorithms'. In: International Conference on Data Mining (ICDM). Sorrento, Italy, 17th Nov. 2020. URL: https://hal.archives-ouvertes.fr/hal-03070560.

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

Scientific book chapters

[34] J. Nicolas. 'Artificial Intelligence and Bioinformatics'. In: *A Guided Tour of Artificial Intelligence Research*. Vol. III. Interfaces and Applications of Artificial Intelligence. 2020, p. 575. URL: https://hal.inria.fr/hal-01850570.

Doctoral dissertations and habilitation theses

[35] W. Delage. 'Characterization and detection of large constitutional insertions for medical use'. Université Rennes 1, 11th Dec. 2020. URL: https://tel.archives-ouvertes.fr/tel-03084361.

[36] L. Lecompte. 'Structural variant genotyping with long read data'. Université de Rennes 1 (UR1), 4th Dec. 2020. URL: https://tel.archives-ouvertes.fr/tel-03082460.

Reports & preprints

[37] 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*. 15th Feb. 2021. DOI: 10.1101/2021.02.12.430979. URL: https://hal.inria.fr/hal-03141144.

[38] V. Epain, R. Andonov, H. Djidjev and D. Lavenier. *Long Reads Assembly Using Integer Linear Programming*. 7th Feb. 2020. URL: https://hal.archives-ouvertes.fr/hal-03007132.

[39] D. J. 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*. 9th Feb. 2020. URL: https://hal.inria.fr/hal-02399723.

Other scientific publications

[40] C. Delahaye and J. Nicolas. *Nanopore MinION long read sequencer: an overview of its error landscape*. 23rd Nov. 2020. URL: https://hal.inria.fr/hal-03123133.

[41] V. Epain. 'DNA fragments positioning improvement to realise consensus sequences in the de novo long reads assembly context'. Université de Rennes 1 [UR1], 16th June 2020. URL: https://hal.inria.fr/hal-03119772.

[42] A. Guichard, F. Legeai, A. Le Bars, P. Y. Jay, M. Joron, D. Tagu and C. Lemaitre. *MTG-Link: filling gaps in draft genome assemblies with linked read data*. Virtual, France, 5th Oct. 2020. URL: https://hal.archives-ouvertes.fr/hal-03074227.

[43] A. Guichard, F. Legeai, A. Le Bars, P. Y. Jay, M. Joron, D. Tagu and C. Lemaitre. *MTG-Link: filling gaps in draft genome assemblies with linked read data*. Montpellier, France, 30th June 2020. URL: https://hal.archives-ouvertes.fr/hal-03073966.

[44] G. Siekaniec, E. Roux, E. Guédon and J. Nicolas. *Bacterial strains identification using Oxford Nanopore sequencing*. Montpellier, France, 30th June 2020. URL: https://hal.archives-ouvertes.fr/hal-03121440.

[45] K. Thuillier. 'AlPha: A Mixed Integer Linear Programming Approach for Genome Haplotyping'. University of Rennes 1, 19th Aug. 2020. URL: https://hal.inria.fr/hal-03127775.

12.3 Other

Scientific popularization

[46] J. Nicolas. 'IA et Bioinformatique'. In: *L'intelligence artificielle: De quoi s'agit-il vraiment ?* 1st July 2020, p. 101. URL: https://hal.inria.fr/hal-03127545.