Activity Report 2019

Team GENSCALE

Scalable, Optimized and Parallel Algorithms for Genomics

Joint team with Inria Rennes – Bretagne Atlantique

D7 – Data and Knowledge Management

Table of contents

1. Team, Visitors, External Collaborators **1**
2. Overall Objectives **2**
2.1. Genomic data processing 2
2.2. Life science partnerships 3
3. Research Program **3**
3.1. Axis 1: Data Structures 3
3.2. Axis 2: Algorithms 3
3.3. Axis 3: Parallelism 4
4. Application Domains **4**
4.1. Introduction 4
4.2. Health 4
4.3. Agronomy 5
4.4. Environment 5
5. Highlights of the Year **5**
6. New Software and Platforms **5**
6.1. SVJedi 5
6.2. MinYS 6
6.3. Simka 6
6.4. DiscoSnpRad 6
7. New Results **7**
7.1. Algorithms & Methods 7
7.1.1. SV genotyping 7
7.1.2. Genome assembly of targeted organisms in metagenomic data 7
7.1.3. SimkaMin: subsampling the kmer space for efficient comparative metagenomics 8
7.1.4. Haplotype reconstruction: phasing co-localized variants 8
7.1.5. Finding all maximal perfect haplotype blocks in linear time 8
7.1.6. Short read correction 8
7.1.7. Large-scale kmer indexation 8
7.1.8. Proteogenomics workflow for the expert annotation of eukaryotic genomes 8
7.1.9. Gap-filling with linked-reads data 9
7.1.10. Statistically Significant Discriminative Patterns Searching 9
7.2. Optimisation 9
7.2.1. Chloroplast genome assembly 9
7.2.2. Integer Linear Programming for Metabolic Networks 9
7.2.3. Integer Linear Programming for De novo Long Reads Assembly 10
7.3. Experiments with the MinION Nanopore sequencer 10
7.3.1. Storing information on DNA Molecules 10
7.3.2. Identification of bacterial strains 10
7.4. Benchmarks and Reviews 10
7.4.1. Evaluation of error correction tools for long Reads 10
7.4.2. Evaluation of insertion variant callers on real human data 10
7.4.3. Modeling activities in cooperation with Inria project Dyliss 11
7.5. Bioinformatics Analysis 11
7.5.1. Genomics of Brassicaceae and agro-ecosystems insects 11
7.5.2. Structural genome analysis of S. pyogenes strains 11
7.5.3. Linking allele-specific expression and natural selection in wild populations 12
8. Bilateral Contracts and Grants with Industry **12**
8.1. Bilateral Contracts with Industry 12
8.2. Bilateral Grants with Industry 12

9. Partnerships and Cooperations **12**
9.1. Regional Initiatives 12
9.1.1. Project Thermin: Differential characterization of strains of a bacterial species, Streptococcus thermophilus, with a Nanopore MinION 12
9.1.2. Project DNA-Store: Advanced error correction scheme for DNA-based data storage using nanopore technology 13
9.2. National Initiatives 13
9.2.1. ANR 13
9.2.1.1. Project HydroGen: Metagenomics applied to ocean life study 13
9.2.1.2. Project SpeCrep: speciation processes in butterflies 13
9.2.1.3. Project Supergene: The consequences of supergene evolution. 13
9.2.1.4. Project SeqDigger: Search engine for genomic sequencing data 14
9.2.2. PIA: Programme Investissement d'Avenir 14
9.2.3. Programs from research institutions 14
9.3. European Initiatives 15
9.3.1. Collaborations in European Programs, Except FP7 & H2020 15
9.3.2. Collaborations with Major European Organizations 15
9.4. International Initiatives 15
9.4.1. HipcoGen 15
9.4.2. Inria International Partners 16
9.5. International Research Visitors 16
9.5.1. Visits of International Scientists 16
9.5.2. Visits to International Teams 16
10. Dissemination **16**
10.1. Promoting Scientific Activities 16
10.1.1. Scientific Events: Selection 16
10.1.1.1. Chair of Conference Program Committees 16
10.1.1.2. Member of the Conference Program Committees 16
10.1.1.3. Reviewer 16
10.1.2. Journal 16
10.1.3. Invited Talks 17
10.1.4. Leadership within the Scientific Community 17
10.1.5. Scientific Expertise 17
10.1.6. Research Administration 17
10.2. Teaching - Supervision - Juries 17
10.2.1. Teaching 17
10.2.2. Supervision 18
10.2.3. Juries 18
10.3. Popularization 19
10.3.1. Internal or external Inria responsibilities 19
10.3.2. Interventions 19
10.3.3. Internal action 19
10.3.4. Creation of media or tools for science outreach 19
11. Bibliography **19**

Project-Team GENSCALE

Creation of the Team: 2012 January 01, updated into Project-Team: 2013 January 01

Keywords:

Computer Science and Digital Science:

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)
A3.3.2. - Data mining
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
B3.6. - Ecology
B3.6.1. - Biodiversity

1. Team, Visitors, External Collaborators

Research Scientists

Dominique Lavenier [Team leader, CNRS, Senior Researcher, HDR]
Claire Lemaitre [Inria, Researcher]
Jacques Nicolas [Inria, Senior Researcher, HDR]
Pierre Peterlongo [Inria, Researcher, HDR]

Faculty Members

Roumen Andonov [Univ de Rennes I, Professor, HDR]
Emeline Roux [Univ de Lorraine, Associate Professor]

Post-Doctoral Fellow

Céline Le Beguec [Inria, Post-Doctoral Fellow]

PhD Students

Kévin Da Silva [Inria, PhD Student]
Wesley Delage [Inria, PhD Student]
Clara Delahaye [Univ de Rennes I, PhD Student, from Oct 2019]
Sébastien Francois [Univ de Rennes I, PhD Student, until Sep 2019]
Rati Kar [INRA, PhD Student, until Jan 2019]
Lolita Lecompte [Inria, PhD Student]
Téo Lemane [Inria, PhD Student, from Oct 2019]
Grégoire Siekaniec [INRA, PhD Student]

Technical staff

Charles Deltel [Inria, Engineer]

Anne Guichard [INRA, Engineer, from Jul 2019]
Gwendal Virlet [CNRS, Engineer, until Sep 2019]

Interns and Apprentices
Ariane Badoual [Inria, from Apr 2019 until Jul 2019]
Clara Delahaye [Inria, until Jun 2019]
Victor Epain [Inria, from Apr 2019 until Jul 2019]
Arthur Le Bars [Inria, until Jul 2019]
Téo Lemane [Univ de Rennes I, until Jul 2019]
Corentin Raphalen [Inria, from Apr 2019 until Jul 2019]

Administrative Assistants
Marie Le Roic [Inria, Administrative Assistant, from Dec 2019]
Marie Le Roic [Univ de Rennes I, Administrative Assistant, until Nov 2019]

External Collaborators
Susete Alves Carvalho [INRA]
Fabrice Legeai [INRA]
Mohammed Amin Madoui [CEA, from Sep 2019]

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 hundreds of Giga bytes of raw data absolutely need to be represented in a very concise way in order to 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 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 [10], to detect structural variants using local NGS assembly approaches [9] 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].

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 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. Highlights of the Year

5.1. Highlights of the Year

5.1.1. Awards

The Gilles Kahn *accessits* **prize was awarded to Camille Marchet** for her PhD thesis: *From reads to transcripts: de novo methods for the analysis of transcriptome second and third generation sequencing* [8]. This thesis was prepared in the GenScale team under the supervision of P. Peterlongo.

The Gilles Kahn prize is awarded each year by the SIF, the French Society of Computer Science, for an excellent PhD thesis in the field of computer science.

The thesis of Camille dealt with the processing of transcriptome sequencing data. More precisely, the question was how to take advantage of the characteristics of the data produced by third generation sequencing technologies, as they produce large sequences covering the total length of RNA molecules. The core work of this thesis consisted in the methodological development and implementation of new algorithms allowing the clustering of third generation sequences by gene, then their correction and finally the detection of the different isoforms of each gene.

6. New Software and Platforms

6.1. 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).

- Participants: Claire Lemaitre, Lolita Lecompte, Pierre Peterlongo and Dominique Lavenier
- Contact: Claire Lemaitre
- URL: https://github.com/llecompte/SVJedi

6.2. MinYS

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.

- Contact: Claire Lemaitre
- URL: https://github.com/cguyomar/MinYS

6.3. 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 FUNCTIONAL DESCRIPTION: 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.

- Participants: Claire Lemaitre, Dominique Lavenier, Gaëtan Benoit and Pierre Peterlongo
- Contact: Claire Lemaitre
- URL: https://gatb.inria.fr/software/simka/

6.4. DiscoSnpRad

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

- Participants: Pierre Peterlongo and Claire Lemaitre
- Contact: Pierre Peterlongo
- URL: https://github.com/GATB/DiscoSnp

7. New Results

7.1. Algorithms & Methods

7.1.1. SV 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. Although many SV callers have been published recently, there is no published method to date dedicated to genotyping SVs with this type of data. Variant genotyping consists in estimating the presence and ploidy or absence of a set of known variants in a newly sequenced individual. Thus, in this paper, we present a new 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 [23], [24], [35]. SVJedi is implemented in Python and available at https://github.com/llecompte/SVJedi.

7.1.2. 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 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. MinYS is a Python3 pipeline, distributed on github (https://github.com/cguyomar/MinYS) and as a conda package in the bio-conda repository [22].

7.1.3. SimkaMin: subsampling the kmer space for efficient comparative metagenomics

Participants: Claire Lemaitre, Pierre Peterlongo.

SimkaMin [12] 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 ($\sim$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.

7.1.4. Haplotype reconstruction: phasing co-localized variants

Participants: Mohammed Amin Madoui, Pierre Peterlongo.

In collaboration with Amin Madoui from the Genoscope (CEA), we develop a new methodology to reconstruct haplotypes or strain genomes directly from raw sequencing set of (metagenomic) reads. The goal is to propose long assembled sequences (i.e. complete genomes are not mandatory) such that each assembled sequence belongs to only one sequenced chromosome and is not a consensus of several similar sequences. Downstream, this enables to perform population genomics analyses.

The key idea is to use the DiscoSnp [10] output, detecting set of variant alleles that are co-localized on input reads or pairs of input reads. Then we finally reconstruct set of sequences that are as parsimonious as possible with those observations.

7.1.5. Finding all maximal perfect haplotype blocks in linear time

Participant: Pierre Peterlongo.

Recent large-scale community sequencing efforts allow at an unprecedented level of detail the identification of genomic regions that show signatures of natural selection. However, traditional methods for identifying such regions from individuals' haplotype data require excessive computing times and therefore are not applicable to current datasets. In 2019, Cunha et al. (Proceedings of BSB 2019) suggested the maximal perfect haplotype block as a very simple combinatorial pattern, forming the basis of a new method to perform rapid genome-wide selection scans. The algorithm they presented for identifying these blocks, however, had a worst-case running time quadratic in the genome length. It was posed as an open problem whether an optimal, linear-time algorithm exists. We gave two algorithms that achieve this time bound, one which is conceptually very simple and uses suffix trees and a second one using the positional Burrows-Wheeler Transform, that is very efficient also in practice [20].

7.1.6. Short read correction

Participant: Pierre Peterlongo.

We propose a new approach for the correction of NGS reads. This approach is based on the construction of a clean de Bruijn graph in which the correction is made at the contig level. In a second step, original reads are mapped on this graph, allowing to correct the original reads [16].

7.1.7. Large-scale kmer indexation

Participants: Téo Lemane, Pierre Peterlongo.

In the SeqDigger ANR project framework (see dedicated Section), we aim to index TB or PB of genomic sequences, assembled or not. The central idea is to assign any kmer (word of length k) to the set of indexed dataset it belongs to. For doing this we have proposed a method that improves one of the state of the art algorithm (HowDeSBT [38]) by optimizing the way kmers are counted and represented [36].

7.1.8. Proteogenomics workflow for the expert annotation of eukaryotic genomes

Participant: Pierre Peterlongo.

Accurate structural annotation of genomes is still a challenge, despite the progress made over the past decade. The prediction of gene structure remains difficult, especially for eukaryotic species, and is often erroneous and incomplete. In [15], we proposed a proteogenomics strategy, taking advantage of the combination of proteomics datasets and bioinformatics tools, to identify novel protein coding-genes and splice isoforms, to assign correct start sites, and to validate predicted exons and genes.

7.1.9. Gap-filling with linked-reads data

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

We develop a novel approach for filling assembly gaps with linked reads data (typically 10X Genomics technology). The approach is based on local assembly using our tool MindTheGap [9], and takes advantage of barcode information to reduce the input read set in order to reduce the de Bruijn graph complexity. The approach is applied to recover the genomic structure of a 1.3 Mb locus of interest in a dozen of re-sequenced butterfly genomes (*H. numata*) in the Supergene ANR project context.

7.1.10. Statistically Significant Discriminative Patterns Searching

Participants: Gwendal Virlet, Dominique Lavenier.

We propose a novel algorithm, called SSDPS, to discover patterns in two-class datasets. The algorithm, developed in collaboration with the LACODAM Inria team, owes its efficiency to an original enumeration strategy of the patterns, which allows to exploit some degrees of anti-monotonicity on the measures of discriminance and statistical significance. Experimental results demonstrate that the performance of the algorithm is better than others. In addition, the number of generated patterns is much less than the number of the other algorithms. An experiment on real data also shows that SSDPS efficiently detects multiple SNPs combinations in genetic data [27].

7.2. Optimisation

7.2.1. Chloroplast genome assembly

Participants: Sébastien Francois, Roumen Andonov, Dominique Lavenier.

This research focuses on the last two stages of *de novo* genome assembly, namely, scaffolding and gap-filling, and shows that they can be solved as part of a single optimization problem. Our approach is based on modeling genome assembly as a problem of finding a simple path in a specific graph that satisfies as many distance constraints as possible encoding the read-pair insert-size information. We formulate it as a mixed-integer linear programming (MILP) problem and apply an optimization solver to find the exact solutions on a benchmark of chloroplast genomes. We show that the presence of repetitions in the set of unitigs is the main reason for the existence of multiple equivalent solutions that are associated to alternative subpaths. We also describe two sufficient conditions and we design efficient algorithms for identifying these subpaths. Comparisons of the results achieved by our tool with the ones obtained with recent assemblers are also presented [11].

7.2.2. Integer Linear Programming for Metabolic Networks

Participants: Kerian Thuillier, Roumen Andonov.

Metabolic networks are a helpful tool to represent and study cell metabolisms. They contain information about every reaction occurring inside an organism. However, metabolic networks of poorly studied species are often incomplete. It is possible to complete these networks with knowledge of other well-known species.

In this study, we present a new linear programming approach for the problem of topological activation in metabolic networks based on flows and the Miller, Tucker and Zemlin (MTZ) formulation for solving the longest path problem. We developed a tool called *Flutampl* with AMPL (A Mathematical Programming Language). It returns optimal solutions for the hybrid completion directly from *sbml* files (the data format used for modelling metabolic networks) [37].

7.2.3. Integer Linear Programming for De novo Long Reads Assembly

Participants: Victor Epain, Roumen Andonov, Dominique Lavenier.

To tackle the de novo long read assembly problem, we investigate a new 2-step method based on integer linear programming. The first step orders the long reads and the second one generates a consensus sequence. Each step is based on a different IPL specification. In 2019, we focused on step 1: long reads are first compared to build an overlapping graph. Then we use integer linear programming to find the heaviest path in a graph $G = (V, E, \lambda)$, where V is the vertices set corresponding to the long reads, E the edge set associated to the overlaps between long reads and λ the overlap length. For large graph, V is partitioned into several parts, each one is solved independently, and the solutions are merged together. Preliminary experimentation show that bacteria assemblies can be successfully solved in a few minutes [31].

7.3. Experiments with the MinION Nanopore sequencer

7.3.1. Storing information on DNA Molecules

Participants: Dominique Lavenier, Emeline Roux, Ariane Badoual.

In 2019, we started a new research activity aiming to explore the possibility to use the DNA molecules as a storage medium. We designed a complete DNA storage system based on the Oxford Nanopore sequencing technology and performed several experimentations by sequencing several synthetized DNA fragments ranging from 500 to 1,000 bp. These sequences have been designed with ad-hoc coding to prevent specific sequencing errors of the nanopore technology such as indel errors in homo-polymer sequences [29] [34]. These real experiments demonstrate that a text encoded into the DNA alphabet, then synthetized into DNA molecules, sequenced with the MinION, and finally processed using bioinformatics approaches can be fully recovered [28].

7.3.2. Identification of bacterial strains

Participants: Jacques Nicolas, Emeline Roux, Grégoire Siekaniec, Clara Delahaye.

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 INRA STLO, which have been made available on the Microscope platform at Genoscope (http://www.genoscope.cns.fr/agc/microscope/). We have developed a first classifier that demonstrates the possibility to identify isolated strains with spaced seed indexing of the noisy long reads produced by the MinIon.

7.4. Benchmarks and Reviews

7.4.1. Evaluation of error correction tools for long Reads

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 [17]. ELECTOR is freely available at https://github.com/kamimrcht/ELECTOR.

7.4.2. 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 reads 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 [9] which overtook others on simulated data, can reach at most 5 to 10 % of the referenced insertion variants. In this work, we propose a classification of the different types of insertion variants, based on the genomic context of the insertion site and the levels of duplication contained in the inserted sequence or within its breakpoints. In a detailed benchmark, we then analyze which of these types are the most impacted by the low recall of existing methods. Finally, by simulating various identified factors of difficulty, we investigate the causes of low recall and how these can be bypassed or improved in existing algorithms.

7.4.3. Modeling activities in cooperation with Inria project Dyliss

Participant: Jacques Nicolas.

J. Nicolas has maintained a partial activity with its previous research team Dyliss. In this framework, we have explored the use of Formal Concept Analysis (FCA) to ease the analysis of biological networks. The PhD thesis of L. Bourneuf on graph compression using FCA, defended this year, has introduced a new extension of FCA for this purpose, working on triplet concepts, which correspond to overlapping bicliques in graphs. The search space of concepts for graph compression has been presented in [21]. FCA applied to data on the steady states of a Boolean network and the dependencies between its proteins allowed to build a classifier used to analyze the states according to the phenotypic signatures of its network components. We have identified variants to the phenotypes and characterized hybrid phenotypes [19].

7.5. Bioinformatics Analysis

7.5.1. Genomics of Brassicaceae and agro-ecosystems insects

Participant: Fabrice Legeai.

Through its long term collaboration with INRA 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. First, on plant genomics, we helped to identify duplicated copies of genes and repeated elements in the Brassica genomes [14]. Then, on major agricultural pests or their natural ennemies such as parasitoids, we conducted large scale analyses on the expression of effector genes involved in the adaptation of pea aphids to their host-plants [13]. Finally, we explored the expression of genes related to the virus machinery of bathyplectes parasitoids wasp of the alfalfa weevil [18].

7.5.2. Structural genome analysis of S. pyogenes strains

Participants: Emeline Roux, Dominique Lavenier.

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 variation.

7.5.3. Linking allele-specific expression and natural selection in wild populations

Participants: Mohammed Amin Madoui, Pierre Peterlongo.

Allele-specific expression (ASE) is now a widely studied mechanism at cell, tissue and organism levels. However, population-level ASE and its evolutionary impacts have still never been investigated. Here, we hypothesized a potential link between ASE and natural selection on the cosmopolitan copepod *Oithona similis*. We combined metagenomic and metatranscriptomic data from seven wild populations of the marine copepod *O. similis* sampled during the Tara Oceans expedition. We detected 587 single nucleotide variants (SNVs) under ASE and found a significant amount of 152 SNVs under ASE in at least one population and under selection across all the populations. This constitutes a first evidence that selection and ASE target more common loci than expected by chance, raising new questions about the nature of the evolutionary links between the two mechanisms [33].

8. Bilateral Contracts and Grants with Industry

8.1. Bilateral Contracts with Industry

8.1.1. Tank milk analysis

Participants: Dominique Lavenier, Jacques Nicolas.

The Seenergi company has developed a biotechnology protocol to detect cow mastitis directly by analyzing the DNA in the milk of the tanks. Cows are first genotyped. Since cows with mastitis produce a high level of lymphocytes, a DNA milk analysis can point out infested cows. Currently, DNA chips are used to support this analysis. We are currently investigating the possibility to use sequencing technologies in order to both reduce cost analysis and to extend the detection to larger herds.

8.2. Bilateral Grants with Industry

8.2.1. Rapsodyn project

Participants: Dominique Lavenier, Claire Lemaitre, Pierre Peterlongo, Gwendal Virlet.

RAPSODYN is a long term project funded by the IA ANR French program (Investissement d'Avenir) and several field seed companies, such as Biogemma, Limagrain and Euralis (http://www.rapsodyn.fr/). The objective 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 analysis.

9. Partnerships and Cooperations

9.1. Regional Initiatives

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

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

Coordinator: J. Nicolas (Inria/Irisa, GenScale, Rennes)
Duration: 36 months (Oct. 2018 – Sept. 2021)
Partners: INRA (STLO, Agrocampus Rennes, E. Guédon and Y. Le Loir).

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 (INRA -Inria collaboration, INRA 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.

9.1.2. 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)

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.

9.2. National Initiatives

9.2.1. ANR

9.2.1.1. Project HydroGen: Metagenomics applied to ocean life study

Participants: Dominique Lavenier, Pierre Peterlongo, Claire Lemaitre.

Coordinator: P. Peterlongo (Inria/Irisa, GenScale, Rennes)
Duration: 42 months (Nov. 2014 – Apr. 2019)
Partners: CEA (Genoscope, Evry), INRA (AgroParisTech, Paris – MIG, Jouy-en-Jossas).

The HydroGen project aims to design new statistical and computational tools to measure and analyze biodiversity through comparative metagenomic approaches. The support application is the study of ocean biodiversity based on the analysis of seawater samples generated by the Tara Oceans expedition.

9.2.1.2. Project SpeCrep: speciation processes in butterflies

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

Coordinator: M. Elias (Museum National d'Histoire Naturelle, Institut de Systématique et d'Evolution de la Biodiversité, Paris)
Duration: 48 months (Jan. 2015 – Jul. 2019)
Partners: MNHN (Paris), INRA (Versailles-Grignon), Genscale Inria/IRISA Rennes.

The SpeCrep project aims at better understanding the speciation processes, in particular by comparing natural replicates from several butterfly species in a suture zone system. GenScale's task is to develop new efficient methods for the assembly of reference genomes and the evaluation of the genetic diversity in several butterfly populations.

9.2.1.3. Project Supergene: The consequences of supergene evolution.

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

Coordinator: M. Joron (Centre d'Ecologie Fonctionnelle et Evolutive (CEFE) UMR CNRS 5175, Montpellier)
Duration: 48 months (Nov. 2018 – Oct. 2022)
Partners: CEFE (Montpellier), MNHN (Paris), Genscale Inria/IRISA Rennes.

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.

9.2.1.4. Project SeqDigger: Search engine for genomic sequencing data

Participants: Dominique Lavenier, Claire Lemaitre, Pierre Peterlongo.

Coordinator: P. Peterlongo
Duration: 48 months (jan. 2020 – Dec. 2024)
Partners: Genscale Inria/IRISA Rennes, CEA genoscpoe, MIO Marseille, Institut Pasteur Paris

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

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.

9.2.2. PIA: Programme Investissement d'Avenir

9.2.2.1. RAPSODYN: Optimization of the rapeseed oil content under low nitrogen

Participants: Dominique Lavenier, Claire Lemaitre, Pierre Peterlongo, Gwendal Virlet.

Coordinator: N. Nesi (Inra, IGEPP, Rennes)
Duration: 99 months (2012-2020)
Partners: 5 companies, 9 academic research labs.

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. (http://www.rapsodyn.fr)

9.2.3. Programs from research institutions

9.2.3.1. Inria Project Lab: Neuromarkers

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

Coordinator: O. Colliot (Inria, Aramis, Paris)
Duration: 4 years (2017-2020)
Partners: Inria (Aramis, Pasteur, Dyliss, GenScale, XPOP), ICM

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.

9.3. European Initiatives

9.3.1. Collaborations in European Programs, Except FP7 & H2020

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.

9.3.2. Collaborations with Major European Organizations

Partner : PHC RILA 2019, Bulgaria

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.

9.4. International Initiatives

9.4.1. HipcoGen

Title: High-Performance Combinatorial Optimization for Computational Genomics

International Partner (Institution - Laboratory - Researcher):

> Information Sciences group of Los Alamos National Laboratory (LANL), Los Alamos, NM 87544, USA. coordinator - Hristo Djidjev

Start year: 2017

See also: https://team.inria.fr/genscale/presentation/associated-team/

Genome sequencing and assembly, the determination of the DNA sequences of a genome, is a core experiment in computational biology. During the last decade, the cost of sequencing has decreased dramatically and a huge amount of new genomes have been sequenced. Nevertheless, most of recent genome projects stay unfinished and nowadays the databases contain much more incompletely assembled genomes than whole stable reference genomes. The main reason is that producing a complete genome, or an as-complete-as-possible-genome, is an extremely difficult computational task (an NP-hard problem) and, in spite of the efforts and the progress done by the bioinformatics community, no satisfactory solution is available today. New sequencing technologies (such as PacBio

or Oxford Nanopore) are being developed that tend to produce longer DNA sequences and offer new opportunities, but also bring significant new challenges. The goal of this joint project, a cooperation between Los Alamos National Laboratory, US and Inria, is to develop a new methodology and tools based on novel optimization techniques and massive parallelism suited to these emerging technologies and able to tackle the complete assembly of large genomes.

9.4.2. Inria International Partners

9.4.2.1. Informal International Partners

- Free University of Brussels, Belgium: Genome assembly [P. Perterlongo, D. Lavenier]

9.5. International Research Visitors

9.5.1. Visits of International Scientists

- Visit of Hristo Djidjev from Los Alamos National Laboratory, USA, June 2019
- Visit of Alla Lapidus and Anto korobeynikov, Center for Algoritmic Biotechnology, St. Petersburg State University, Russia, October 2019

9.5.2. Visits to International Teams

9.5.2.1. Research Stays Abroad

- Visit of R. Andonov at Los Alamos National Laboratory, USA, from March 23 to April 30th, 2019.
- Visit of D. Lavenier at Los Alamos National Laboratory, USA, from May 13th to May 24th, 2019

10. Dissemination

10.1. Promoting Scientific Activities

10.1.1. Scientific Events: Selection

10.1.1.1. Chair of Conference Program Committees

- seqBIM 20019 : national meeting of the sequence algorithms GT seqBIM [C. Lemaitre]

10.1.1.2. Member of the Conference Program Committees

- JOBIM 2019: French symposium of Bioinformatics [C. Lemaitre]
- seqBIM 2019 : national meeting of the sequence algorithms GT seqBIM [C. Lemaitre, P. Peterlongo]
- BIBM 2019, IEEE International Conference on Bioinformatics and Biomedicine [D. Lavenier]
- BIOKDD 2019, 10th International Workshop on Biological Knowledge Discovery from Big Data [D. Lavenier]
- IPDPS 2019, 33rd IEEE International Parallel and Distributed Processing Symposium [D. Lavenier]
- 11th International Conference on Bioinformatics and Computational Biology (BICOB-2019) [R. Andonov]
- RCAM 2019 (Recent Computational Advances in Metagenomics) [P. Peterlongo]

10.1.1.3. Reviewer

- ISMB 2019 (Intelligent Systems for Molecular Biology) [W. Delage, C. Lemaitre]

10.1.2. Journal

10.1.2.1. Reviewer - Reviewing Activities

- Bioinformatics [P. Peterlongo, D. Lavenier]
- BMC Bioinformatics [D. Lavenier]

- BMC Genomics [D. Lavenier, F. Legeai]
- Briefing in Bioinformatics [D. Lavenier]
- Nature Scientific Reports [F. Legeai]
- Genome Biology [D. Lavenier]
- Genomics [D. Lavenier]
- Machine Learning [J. Nicolas]

10.1.3. Invited Talks

- (IB)2 Research Day - Bruxelles - Oct 2019- "Metagenomics-Comparisons and variants" [P. Peterlongo]

10.1.4. Leadership within the Scientific Community

- Members of the Scientific Advisory Board of the GDR BIM (National Research Group in Molecular Bioinformatics) [P. Peterlongo, C. Lemaitre]
- Animator of the Sequence Algorithms axis (seqBIM GT) of the BIM and IM GDRs (National Research Groups in Molecular Bioinformatics and Informatics and Mathematics respectively) [C. Lemaitre]
- Animator of the INRA Center for Computerized Information Treatment "BBRIC" [F. Legeai]

10.1.5. 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 the Computational Biology Institute of Montpellier [D. Lavenier]
- Member of the Scientific Council of Agrocampus Ouest (Institute for life, food and horticultural sciences and landscaping) [J. Nicolas]
- Expert for the Elixir French Service Delivery Plan [P. Peterlongo]

10.1.6. 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]
- Member of the steering committee of the INRA BIPAA Platform (BioInformatics Platform for Agro-ecosystems Arthropods) [D. Lavenier]
- Member of the steering committee of The GenOuest Platform (Bioinformatics Resource Center BioGenOuest) [D. Lavenier]
- Scientific Advisor of The GenOuest Platform (Bioinformatics Resource Center BioGenOuest) [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 (90 students) [R. Andonov]
- Member of the Council of Administration of ISTIC [R. Andonov]

10.2. Teaching - Supervision - Juries

10.2.1. Teaching

Licence : C. Delahaye, G. Siekaniec, Python, 48 h, L1, Univ. Rennes 1, France.

Licence : R. Andonov, Graph Algorithms, 60h, L3, Univ. Rennes 1, France.

Licence : K. Da Silva, Algorithms and Complexity, 36 h, L3, ENSAI, Rennes, France.

Licence : W. Delage, Introduction to Biostatistics, 36 h, L2, Univ. Rennes 1, France.

Master : R. Andonov, S. Francois, Operational research, 82h, M1 Miage, Univ. Rennes 1, France.

Master : G. Siekaniec, Python, 21 h, M1, Univ. Rennes 1, France.

Master : K. Da Silva, Statistical learning, 22h, M1, Univ. Rennes 1, France.

Master : 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, 12h, 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, Engineering School, ESIR, Rennes.

Master : W. Delage, Internship jury, 6 h, M1, Univ. Rennes 1, France

10.2.2. Supervision

PhD in progress : S. François, Combinatorial Optimization Approaches for Bioinformatics, 01/10/2016-30/09/2019, R. Andonov.

PhD in progress : L. Lecompte, Structural Variant detection in long-read sequencing data, 01/09/2017, D. Lavenier and C. Lemaitre.

PhD in progress : W. Delage, De novo local assembly approaches for the detection of complex genomic variations in rare diseases, 01/10/2017, J. Thévenon and C. Lemaitre.

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

10.2.3. Juries

- *Member of Habilitation thesis jury*. S. Caboche, Univ. Lille [P. Peterlongo] R. Eyraud, Univ. Marseille [J. Nicolas, referee]
- *President of PhD thesis jury*. J. Wanza, University of Nice Sophia Antipolis [D. Lavenier], M. Guillemé, Univ. Rennes [J. Nicolas]
- *Referee of Ph-D thesis jury*. Kevin Gravouil, Univ Clermont Auvergne [P. Peterlongo]
- *Member of PhD thesis juries*. Lyam Baudry, University Paris-Sorbonne [C. Lemaitre], Pierre Morisse, Univ. Rouen [P. Peterlongo], Pierre Marijon, Univ Lille [P. Peterlongo], Franklin Delehelle, University of Toulouse [D. Lavenier], Lucas Bourneuf, Univ. Rennes [J. Nicolas].
- *Member of PhD thesis committee*. Chi Nguyen Lam, Univ. Brest [D. Lavenier], Benjamin Churcheward, Univ. Nantes [D. Lavenier], Victor Gaborit, Inserm Nantes [P. Peterlongo], Guillaume Gautreau CEA [P. Peterlongo], Sébastien François, Univ. Rennes [J. Nicolas], Mikaïl Demirdelen, Univ. Rennes [J. Nicolas], Hugo Talibart, Univ. Rennes [J. Nicolas].

10.3. Popularization

10.3.1. Internal or external Inria responsibilities

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

10.3.2. Interventions

- In educational institutions : Participation to operation "A la découverte de la recherche" in high schools [P. Peterlongo]
- Short Movie "Moi, mon chien et les autres: à la recherche des transcrits perdus", presented at Sciences en Courts, a local contest of popularization short movies made by PhD students (http://sciences-en-courts.fr/) [G. Siekaniec].

10.3.3. Internal action

- co-organizer of Sciences en cour[t]s event, Nicomaque association (http://sciences-en-courts.fr/) [W. Delage]

10.3.4. Creation of media or tools for science outreach

- Popularization article in Interstices. “Analyser le génome des océans” https://interstices.info/analyser-les-genomes-des-oceans/ [P. Peterlongo]

11. Bibliography

Major publications by the team in recent years

[1] G. BENOIT, C. LEMAITRE, D. LAVENIER, E. DREZEN, T. DAYRIS, R. URICARU, G. RIZK. *Reference-free compression of high throughput sequencing data with a probabilistic de Bruijn graph*, in "BMC Bioinformatics", September 2015, vol. 16, n° 1 [*DOI :* 10.1186/s12859-015-0709-7], https://hal.inria.fr/hal-01214682

[2] G. BENOIT, P. PETERLONGO, M. MARIADASSOU, E. DREZEN, S. SCHBATH, D. LAVENIER, C. LEMAITRE. *Multiple comparative metagenomics using multiset k -mer counting*, in "PeerJ Computer Science", November 2016, vol. 2 [*DOI :* 10.7717/PEERJ-CS.94], https://hal.inria.fr/hal-01397150

[3] R. CHIKHI, G. RIZK. *Space-efficient and exact de Bruijn graph representation based on a Bloom filter*, in "Algorithms for Molecular Biology", 2013, vol. 8, n° 1, 22 p. [*DOI :* 10.1186/1748-7188-8-22], http://hal.inria.fr/hal-00868805

[4] E. DREZEN, G. RIZK, R. CHIKHI, C. DELTEL, C. LEMAITRE, P. PETERLONGO, D. LAVENIER. *GATB: Genome Assembly & Analysis Tool Box*, in "Bioinformatics", 2014, vol. 30, pp. 2959-2961 [*DOI :* 10.1093/BIOINFORMATICS/BTU406], https://hal.archives-ouvertes.fr/hal-01088571

[5] S. FRANÇOIS, R. ANDONOV, D. LAVENIER, 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, March 2018, pp. 1-11 [*DOI :* 10.1101/231324], https://hal.inria.fr/hal-01666830

[6] C. GUYOMAR, F. LEGEAI, E. JOUSSELIN, C. C. MOUGEL, C. LEMAITRE, J.-C. SIMON. *Multi-scale characterization of symbiont diversity in the pea aphid complex through metagenomic approaches*, in "Microbiome", December 2018, vol. 6, n° 1 [*DOI :* 10.1186/s40168-018-0562-9], https://hal.archives-ouvertes.fr/hal-01926402

[7] A. LIMASSET, G. RIZK, R. CHIKHI, P. PETERLONGO. *Fast and scalable minimal perfect hashing for massive key sets*, in "16th International Symposium on Experimental Algorithms", London, United Kingdom, June 2017, vol. 11, pp. 1-11, https://hal.inria.fr/hal-01566246

[8] C. MARCHET. *From reads to transcripts: de novo methods for the analysis of transcriptome second and third generation sequencing*, Université de Rennes 1, September 2018, https://tel.archives-ouvertes.fr/tel-01939193

[9] G. RIZK, A. GOUIN, R. CHIKHI, C. LEMAITRE. *MindTheGap: integrated detection and assembly of short and long insertions*, in "Bioinformatics", December 2014, vol. 30, n° 24, pp. 3451-3457 [*DOI :* 10.1093/BIOINFORMATICS/BTU545], https://hal.inria.fr/hal-01081089

[10] R. URICARU, G. RIZK, V. LACROIX, E. QUILLERY, O. PLANTARD, R. CHIKHI, C. LEMAITRE, P. PETERLONGO. *Reference-free detection of isolated SNPs*, in "Nucleic Acids Research", November 2014, pp. 1-12 [*DOI :* 10.1093/NAR/GKU1187], https://hal.inria.fr/hal-01083715

Publications of the year

Articles in International Peer-Reviewed Journals

[11] R. ANDONOV, H. DJIDJEV, S. FRANÇOIS, D. LAVENIER. *Complete Assembly of Circular and Chloroplast Genomes Based on Global Optimization*, in "Journal of Bioinformatics and Computational Biology", 2019, pp. 1-28, forthcoming [*DOI :* 10.1142/S0219720019500148], https://hal.archives-ouvertes.fr/hal-02151798

[12] G. BENOIT, M. MARIADASSOU, S. ROBIN, S. SCHBATH, P. PETERLONGO, C. LEMAITRE. *SimkaMin: fast and resource frugal de novo comparative metagenomics*, in "Bioinformatics", September 2019, pp. 1-2 [*DOI :* 10.1093/BIOINFORMATICS/BTZ685], https://hal.inria.fr/hal-02308101

[13] H. BOULAIN, F. LEGEAI, J. JAQUIÉRY, E. GUY, S. MORLIERE, J.-C. SIMON, A. SUGIO. *Differential Expression of Candidate Salivary Effector Genes in Pea Aphid Biotypes With Distinct Host Plant Specificity*, in "Frontiers in Plant Science", October 2019, vol. 10, pp. 1-12 [*DOI :* 10.3389/FPLS.2019.01301], https://hal.inria.fr/hal-02378266

[14] J. FERREIRA DE CARVALHO, J. LUCAS, G. DENIOT, C. FALENTIN, O. FILANGI, M. GILET, F. LEGEAI, M. LODE, J. MORICE, G. TROTOUX, J.-M. AURY, V. BARBE, J. KELLER, R. SNOWDON, Z. HE, F. DENOEUD, P. WINCKER, I. BANCROFT, A.-M. CHÈVRE, M. ROUSSEAU-GUEUTIN. *Cytonuclear interactions remain stable during allopolyploid evolution despite repeated whole-genome duplications in Brassica*, in "Plant Journal", February 2019, vol. 98, n° 3, pp. 434-447 [*DOI :* 10.1111/TPJ.14228], https://hal-univ-rennes1.archives-ouvertes.fr/hal-02019346

[15] L. GUILLOT, L. DELAGE, A. VIARI, Y. VANDENBROUCK, E. COM, A. A. RITTER, R. LAVIGNE, D. MARIE, P. PETERLONGO, P. POTIN, C. PINEAU. *Peptimapper: proteogenomics workflow for the expert annotation of eukaryotic genomes*, in "BMC Genomics", January 2019, vol. 20, n° 1, 56 p. [*DOI :* 10.1186/S12864-019-5431-9], https://hal.inria.fr/hal-01987197

[16] A. LIMASSET, J.-F. FLOT, P. PETERLONGO. *Toward perfect reads: self-correction of short reads via mapping on de Bruijn graphs*, in "Bioinformatics", February 2019 [*DOI :* 10.1093/BIOINFORMATICS/BTZ102], https://hal.inria.fr/hal-02407243

[17] C. MARCHET, P. MORISSE, L. LECOMPTE, A. LEFEBVRE, T. LECROQ, P. PETERLONGO, A. LIMASSET. *ELECTOR: evaluator for long reads correction methods*, in "NAR Genomics and Bioinformatics", March 2020, vol. 2, n° 1 [*DOI :* 10.1093/NARGAB/LQZ015], https://hal.inria.fr/hal-02371117

[18] S. ROBIN, M. RAVALLEC, M. FRAYSSINET, J. WHITFIELD, V. JOUAN, F. LEGEAI, A.-N. VOLKOFF. *Evidence for an ichnovirus machinery in parasitoids of coleopteran larvae*, in "Virus Research", 2019, vol. 263, pp. 189-206 [*DOI :* 10.1016/J.VIRUSRES.2019.02.001], https://hal.archives-ouvertes.fr/hal-02059774

[19] M. WERY, O. DAMERON, J. NICOLAS, E. RÉMY, A. SIEGEL. *Formalizing and enriching phenotype signatures using Boolean networks*, in "Journal of Theoretical Biology", 2019, vol. 467, pp. 66-79 [*DOI :* 10.1016/J.JTBI.2019.01.015], https://hal.inria.fr/hal-02018724

International Conferences with Proceedings

[20] J. N. ALANKO, H. BANNAI, B. CAZAUX, P. PETERLONGO, J. STOYE. *Finding all maximal perfect haplotype blocks in linear time*, in "WABI 2019 - Workshop on Algorithms in Bioinformatics", Niagara Falls, United States, ACM, September 2019, pp. 1-9, https://hal.inria.fr/hal-02187246

[21] L. BOURNEUF, J. NICOLAS. *Concept Lattices as a Search Space for Graph Compression*, in "ICFCA 2019 - 15th International Conference on Formal Concept Analysis", Francfort, Germany, D. C. L. B. SERTKAYA (editor), ICFCA: International Conference on Formal Concept Analysis, Springer, May 2019, vol. 15th International Conference, n° 15, pp. 274-289 [*DOI :* 10.1007/978-3-030-21462-3_18], https://hal.inria.fr/hal-02399578

National Conferences with Proceedings

[22] C. GUYOMAR, W. DELAGE, F. LEGEAI, C. MOUGEL, J.-C. SIMON, C. LEMAITRE. *Reference-guided genome assembly in metagenomic samples*, in "JOBIM 2019 - Journées Ouvertes Biologie, Informatique et Mathématiques", Nantes, France, July 2019, pp. 1-8, https://hal.inria.fr/hal-02308257

[23] L. LECOMPTE, P. PETERLONGO, D. LAVENIER, C. LEMAITRE. *Genotyping Structural Variations using Long Read data*, in "JOBIM 2019 - Journées Ouvertes Biologie, Informatique et Mathématiques", Nantes, France, July 2019, pp. 1-8, https://hal.inria.fr/hal-02288091

Conferences without Proceedings

[24] L. LECOMPTE, P. PETERLONGO, D. LAVENIER, C. LEMAITRE. *Genotyping Structural Variations using Long Read Data*, in "HiTSeq 2019 - Conference on High Throughput Sequencing", Basel, Switzerland, July 2019, pp. 1-3, https://hal.inria.fr/hal-02289484

[25] C. MARCHET, M. KERBIRIOU, A. LIMASSET. *Indexing De Bruijn graphs with minimizers*, in "Recomb seq", Whashinton, United States, April 2019 [*DOI :* 10.1101/546309], https://hal.archives-ouvertes.fr/hal-02435086

[26] P. MORISSE, C. MARCHET, A. LIMASSET, T. LECROQ, A. LEFEBVRE. *CONSENT: Scalable self-correction of long reads with multiple sequence alignment*, in "Recomb Seq", Washinton, France, May 2019, https://hal.archives-ouvertes.fr/hal-02435116

[27] H. S. PHAM, G. VIRLET, D. LAVENIER, A. TERMIER. *Statistically Significant Discriminative Patterns Searching*, in "DaWaK 2019 - 21st International Conference on Big Data Analytics and Knowledge Discov-

ery", Linz, Austria, Springer, August 2019, pp. 105-115 [*DOI :* 10.1007/978-3-030-27520-4_8], https://hal.archives-ouvertes.fr/hal-02190793

Other Publications

[28] A. BADUAL. *Stockage d'information sur ADN*, Université Rennes 1, June 2019, https://hal.inria.fr/hal-02401641

[29] L. CONDE-CANENCIA, B. HAMOUM, D. LAVENIER, E. ROUX. *Error Correction Schemes for DNA Storage with Nanopore Sequencing*, July 2019, JOBIM 2019 - Journées Ouvertes Biologie, Informatique et Mathématiques, Poster, https://hal.archives-ouvertes.fr/hal-02400744

[30] K. DA SILVA, N. PONS, M. BERLAND, F. PLAZA OÑATE, M. ALMEIDA, P. PETERLONGO. *From genomics to metagenomics: benchmark of variation graphs*, July 2019, JOBIM 2019 - Journées Ouvertes Biologie, Informatique et Mathématiques, Poster, https://hal.inria.fr/hal-02284559

[31] V. EPAIN. *De novo long reads assembly using integer linear programing*, Université de Rennes 1 [UR1], September 2019, https://hal.inria.fr/hal-02413832

[32] J. GAUTHIER, C. MOUDEN, T. SUCHAN, N. ALVAREZ, N. ARRIGO, C. RIOU, C. LEMAITRE, P. PETERLONGO. *DiscoSnp-RAD: de novo detection of small variants for population genomics*, October 2019, working paper or preprint, https://hal.inria.fr/hal-01634232

[33] R. LASO-JADART, K. SUGIER, E. PETIT, K. LABADIE, P. PETERLONGO, C. AMBROISE, P. WINCKER, J.-L. JAMET, M.-A. MADOUI. *Linking Allele-Specific Expression And Natural Selection In Wild Populations*, September 2019, working paper or preprint [*DOI :* 10.1101/599076], https://hal.inria.fr/hal-02275928

[34] D. LAVENIER, E. ROUX, L. CONDE-CANENCIA, B. HAMOUM. *Advanced Coding Schemes for DNA-Based Data Storage Using Nanopore Sequencing Technologies*, November 2019, Journées CominLabs 2019, Poster, https://hal.archives-ouvertes.fr/hal-02400656

[35] L. LECOMPTE, P. PETERLONGO, D. LAVENIER, C. LEMAITRE. *SVJedi : Structural variation genotyping using long reads*, July 2019, HiTSeq 2019 - Conference on High Throughput Sequencing, Poster, https://hal.inria.fr/hal-02290884

[36] T. LEMANE. *Search engine for genomic sequencing data*, Université Rennes1, July 2019, https://hal.inria.fr/hal-02410102

[37] K. THUILLIER. *Linear programming for metabolic network completion*, Inria Rennes - Bretagne Atlantique and University of Rennes 1, France, December 2019, https://hal.inria.fr/hal-02408003

References in notes

[38] R. S. HARRIS, P. MEDVEDEV. *Improved representation of sequence bloom trees*, in "Bioinformatics", 08 2019, btz662, https://doi.org/10.1093/bioinformatics/btz662