Activity Report 2018

Team GENSCALE

Scalable, Optimized and Parallel Algorithms for Genomics

Joint team with Inria Rennes – Bretagne Atlantique

D7 – Data and Knowledge Management

Table of contents

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, from Jun 2018, HDR]
Pierre Peterlongo [Inria, Researcher, HDR]

Faculty Member

Roumen Andonov [Univ de Rennes I, Professor, HDR]

Post-Doctoral Fellow

Celine Le Beguec [Inria, from Dec 2018]

PhD Students

Kevin Da Silva [Inria, from Oct 2018]
Wesley Delage [Inria]
Sebastien Francois [Univ de Rennes I]
Cervin Guyomar [Univ de Rennes I, until Sep 2018, Inria, from Oct 2018 until Nov 2018]
Rati Kar [INRA, from Aug 2018]
Lolita Lecompte [Inria]
Camille Marchet [Univ de Rennes I, until Sep 2018]
Gregoire Siekaniec [INRA, from Oct 2018]

Technical staff

Charles Deltel [Inria]
Jeremy Gauthier [Inria, until Jul 2018]

Gwendal Virlet [CNRS, from Oct 2018]
Sebastien Letort [CNRS, until Nov 2018, INRA, from Nov 2018]

Interns

Martinien Adda [CNRS, from May 2018 until Aug 2018]
Maxime Bridoux [Inria, from May 2018 until Jul 2018]
Benjamin Churcheward [Inria, until Jul 2018]
Victor Epain [CNRS, from May 2018 until Jul 2018]
Mohamed Moselhy [Inria, from May 2018 until Aug 2018]
Gregoire Siekaniec [Inria, until Jul 2018]
Gwendal Virlet [CNRS, from Mar 2018 until Aug 2018]

Administrative Assistant

Marie Le Roic [Univ de Rennes I]

External Collaborators

Susete Alves Carvalho [INRA]
Fabrice Legeai [INRA]
Emeline Roux [Univ de Lorraine, from Sep 2018]

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 next-generation sequencing (NGS) technologies that provide 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 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 Structure

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, require high computing resources, and more specifically very important memory configuration. For example, the ABYSS software used 4.3TB of memory to assemble the white spruce genome [36]. The main reason for such memory consumption is that the data structures used in ABYSS are far from optimal (and this is also the case for many assembly software).

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 (Next Generation Sequencing) processing requirements (see next section). 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].

Another research direction of this axis 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 [5].

3.2. Axis 2: Algorithms

The main goal of the GenScale team is to develop optimized tools dedicated to NGS 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 NGS 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 [26].

- **Detection of variants** This is often the first information we want to extract from billions of reads. Variant structures range from SNPs or short indels to large insertions/deletions and long inversions over the chromosomes. We developed original methods to find variants without any reference genome [7]. We also worked on the detection of structural variants using approaches of local assembly [6].
- **Metagenomics** We focussed 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].
- **Genome Wide Association Study (GWAS)** We tackle this problem with algorithms commonly used in data mining. From two cohorts of individuals (case and control) we can exhibit statistically significant *patterns* spaning over full genomes.

In addition, we also proposed new algorithmic solutions for analyzing third generation sequencing data, in order to benefit from their larger read size while taking into account their higher sequencing error rate [16].

3.3. Axis 3: Parallelism

This third axis investigates another lever to increase performances and scalability of NGS 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. Today the bioinformatics analysis is mainly based on the detection of SNPs (Single Nucleotide Polymorphism) from a set of predefined target genes. Tomorrow, due to the decreasing cost of the sequencing process, bioinformatics analysis will scan the complete genome and report all kinds of mutations, including complex mutations such as large insertions or deletions, that could be associated with cancers.

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 anatomical 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 and Environment

Improving plant breeding: such projects aim at 1) identifying favorable alleles at loci contributing to phenotypic variation, 2) characterizing polymorphism at the functional level and 3) 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.

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 [14].

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 their role, for example, in the CO2 sequestration.

5. Highlights of the Year

5.1. Highlights of the Year

Our new tool, Carnac-LR, dedicated to the clustering of third generation sequencing data, has been published in the high impact journal *Nucleic Acid Research* (NAR) [16].

5.1.1. Awards

BEST PAPER AWARD:

[26]
S. FRANCOIS, 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. New Software and Platforms

6.1. GATB-Core

Genome Assembly and Analysis Tool Box
KEYWORDS: Bioinformatics - NGS - Genomics - Genome assembling

FUNCTIONAL DESCRIPTION: The GATB-Core library aims to lighten the design of NGS algorithms. It offers a panel of high-level optimized building blocks to speed-up the development of NGS tools related to genome assembly and/or genome analysis. The underlying data structure is the de Bruijn graph, and the general parallelism model is multithreading. The GATB library targets standard computing resources such as current multicore processor (laptop computer, small server) with a few GB of memory. From high-level API, NGS programming designers can rapidly elaborate their own software based on domain state-of-the-art algorithms and data structures. The GATB-Core library is written in C++.

- Participants: Charles Deltel, Rayan Chikhi, Erwan Drezen, Antoine Limasset, Gaëtan Benoit, Uricaru Raluca, Claire Lemaitre, Dominique Lavenier, Guillaume Rizk, Patrick Durand and Pierre Peterlongo
- Contact: Dominique Lavenier
- URL: http://gatb.inria.fr/

6.2. CARNAC-LR

Clustering coefficient-based Acquisition of RNA Communities in Long Reads
KEYWORDS: Transcriptomics - Clustering - Bioinformatics
FUNCTIONAL DESCRIPTION: Carnac-LR is a clustering method for third generation sequencing data. Used on RNA sequences it retrieves all sequences that relate to a same gene and put them in a cluster. CARNAC-LR is an efficient implementation of a novel clustering algorithm for detecting communities in a graph of reads from Third Generation Sequencing. It is a part of a pipeline that allows to retrieve expressed variants from each gene de novo (without reference genome/transcriptome), for transcriptomic sequencing data.

- Participants: Camille Marchet, Pierre Peterlongo and Jacques Nicolas
- Contact: Camille Marchet
- Publication: De Novo Clustering of Long Reads by Gene from Transcriptomics Data
- URL: https://github.com/kamimrcht/CARNAC

6.3. MindTheGap

KEYWORDS: Bioinformatics - NGS - Genome assembly
FUNCTIONAL DESCRIPTION: MindTheGap is a NGS software that performs local assembly of short reads. It is a structural variant detection tool as well as a genome assembly finishing tool. As a variant caller, it performs detection and assembly of DNA insertion variants in NGS read datasets with respect to a reference genome. It is designed to call insertions of any size, whether they are novel or duplicated, homozygous or heterozygous in the donor genome. Local assembly is performed to recover the inserted sequences from the whole read dataset. The local assembly module can also be used to fill the gaps between a set of input contigs without any a priori on their relative order and orientation, in order to improve a draft genome assembly.

RELEASE FUNCTIONAL DESCRIPTION: Since version 2.1.0, MindTheGap can also be used as a genome assembly finishing tool: it can fill the gaps between a set of input contigs without any a priori on their relative order and orientation. This new feature is available in the Fill module with option -contig.

- Participants: Claire Lemaitre, Guillaume Rizk, Pierre Marijon, Rayan Chikhi, Wesley Delage and Cervin Guyomar
- Contact: Claire Lemaitre
- Publication: MindTheGap: integrated detection and assembly of short and long insertions
- URL: https://gatb.inria.fr/software/mind-the-gap/

6.4. bcool

de Bruijn graph cOrrectiOn from graph aLignment
KEYWORDS: De Bruijn graphs - Reads correction - Short reads - Read mapping

FUNCTIONAL DESCRIPTION: BCool is a method to correct short reads using de Bruijn graphs. BCool includes two steps. As a first step, Bcool constructs a corrected compacted de Bruijn graph from the reads. This graph is then used as a reference and the reads are corrected according to their mapping on the graph.This approach yields a better correction than kmer-spectrum techniques, while being scalable, making it possible to apply it to human-size genomic datasets and beyond. The implementation is open source and available at github.com/Malfoy/BCOOL

- Participants: Antoine Limasset and Pierre Peterlongo
- Partner: Université libre de Bruxelles
- Contact: Pierre Peterlongo
- Publication: Toward perfect reads: self-correction of short reads via mapping on de Bruijn graphs
- URL: http://github.com/Malfoy/BCOOL

7. New Results

7.1. Data Structure

7.1.1. Quasi-dictionary data structure

Participants: Camille Marchet, Lolita Lecompte, Pierre Peterlongo.

Indexing massive data sets is extremely expensive for large scale problems. In many fields, huge amounts of data are currently generated. However, extracting meaningful information from voluminous data sets, such as computing similarity between elements, is far from being trivial. It remains nonetheless a fundamental need. This work proposes a probabilistic data structure based on a minimal perfect hash function for indexing large sets of keys. Our structure out-compete the hash table for construction, query times and for memory usage, in the case of the indexation of a static set. To illustrate the impact of algorithms performances, we provide two applications based on similarity computation 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 lower recall quality results.

The quasi-dictionary, is freely available at https://github.com/pierrepeterlongo/quasi_dictionary. The associated paper has been published in Discrete Applied Mathematics [17].

7.2. Algorithms & Methods

7.2.1. Genome assembly of targeted organisms in metagenomic data

Participants: Wesley Delage, Cervin Guyomar, Fabrice Legeai, Claire Lemaitre.

In this work, we propose a two-step reference-guided assembly method tailored for metagenomic data. 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 modified 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 this method 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 5 times faster than classical *de novo* metagenomics assemblers and it is able to recover large structural variations co-existing in a sample. The modified version of MindTheGap is freely available at http://github.com/GATB/MindTheGap (version $>$ 2.1.0) [31].

7.2.2. De Novo Clustering of Long Reads by Gene from Transcriptomics Data

Participants: Camille Marchet, Lolita Lecompte, Jacques Nicolas, Pierre Peterlongo.

Long-read sequencing currently provides sequences of a few thousand base pairs. It is therefore possible to obtain complete transcripts, offering an unprecedented vision of the cellular transcriptome. However the literature lacks tools for *de novo* clustering of such data, in particular for Oxford Nanopore Technologies reads, because of the inherent high error rate compared to short reads. Our goal is to process reads from whole transcriptome sequencing data accurately and without a reference genome in order to reliably group reads coming from the same gene. This *de novo* approach is therefore particularly suitable for non-model species, but can also serve as a useful pre-processing step to improve read mapping. Our contribution both proposes a new algorithm adapted to clustering of reads by gene and a practical and free access tool that allows to scale the complete processing of eukaryotic transcriptomes. We sequenced a mouse RNA sample using the MinION device. This dataset is used to compare our solution to other algorithms used in the context of biological clustering. We demonstrate that it is the best approach for transcriptomics long reads. When a reference is available to enable mapping, we show that it stands as an alternative method that predicts complementary clusters.

The tool, called CARNAC-LR, is freely available at https://github.com/kamimrcht/CARNAC-LR. This work has been published in Nucleic Acids Research journal [16] and presented in several conferences [33], [28].

7.2.3. Comparison of approaches for finding alternative splicing events in RNA-seq

Participant: Camille Marchet.

In this work we compared an assembly-first and a mapping-first approach to analyze RNA-seq data and find alternative splicing (AS) events. Assembly-first approach enables to identify novel AS events and to detect events in paralog genes that are hard to find using mapping because of multiple equivalent matches. On the other hand, the mapping-first approach is more sensitive and detects AS events in lowly expressed genes, and is also able to find AS events with exons containing transposable elements. In addition we support these results with experimental validation. We showed that in order to extensively study the alternative splicing via RNA-seq data and retrieve the most candidates, both approaches should be led. We provide a pipeline consituted of parallel local *de novo* assembly executed by KisSplice and mapping using a novel mapping workflow called FaRLine [11].

7.2.4. Short read correction

Participant: Pierre Peterlongo.

We proposed a new method to correct short reads using de Bruijn graphs, and we implemented it as a tool called Bcool. As a first step, Bcool constructs a corrected compacted de Bruijn graph from the reads. This graph is then used as a reference and the reads are corrected according to their mapping on the graph. We showed that this approach yields a better correction than kmer-spectrum techniques, while being scalable, making it possible to apply it to human-size genomic datasets and beyond [27].

7.2.5. Long read splitting of heterozygous genomes

Participants: Dominique Lavenier, Maxime Bridoux.

This study aims to directly split long reads of highly heterozygous genomes to help assembly. Long read technologies provide very noisy sequences with many short indel errors. Standard assembly software do not really make difference between heterozygosity and sequencing errors. For highly heterozygous genomes this confusion may lead to misassembly. To separate long reads accordingly to their haplotype, we developed a new k-mer based method. After an alignment step to group similar reads, we build slices of 1 kbp along the multiple alignment containing a representative number of reads. The splitting is done by focusing on k-mers that are absent in one group and not in another one. This in an ongoing work started by the internship of M. Bridoux [29] in the framework of the France Genomique ALPAGA project.

7.3. Optimisation

7.3.1. Distance-Constrained Elementary Path Problem

Participants: Sebastien François, Rumen Andonov.

Given a directed graph $G = (V, E, l)$ with weights $l_e \geq 0$ associated with arcs $e \in E$ and a set of vertex pairs with distances between them (called *distance constraints*), the problem is to find an elementary path in G that satisfies a maximum number of distance constraints. We call it *Distance-Constrained Elementary Path (DCEP)* problem. This problem is motivated by applications in genome assembly. We describe three Mixed Integer Programming (MIP) formulations for this problem and discuss their advantages [25].

7.3.2. Complete Assembly of Circular Genomes Based on Global Optimization

Participants: Sebastien François, Rumen Andonov, Dominique Lavenier.

The goal here is to develop a new methodology and tools based on strong mathematical foundations and novel optimization techniques for solving the genome assembly problem. During the current year we focused on the last two stages of genome assembly, namely scaffolding and gap-filling, and showed that they can be solved as part of a single optimization problem. We obtained this by modeling genome assembly as a problem of finding a simple path in a specific graph that satisfies as many as possible of the distance constraints encoding the insert-size information. We formulated it as a mixed-integer linear programming problem and applied an optimization solver to find the exact solution on a benchmark of chloroplasts. Our tool is called GAT (Genscale Assembly Tool) and we tested it on a set of 33 chloroplast genome data. Comparisons with some of the most popular recent assemblers show that our tool produces assemblies of significantly higher quality than these heuristics [26]. These results fully justify the efforts for designing exact approaches for genome assembly.

7.4. Parallelism

7.4.1. Variant detection using processing-in-memory technology

Participants: Dominique Lavenier, Mohamed Moselhy.

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. In 2018, we define 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. Performance evaluation on the FPGA UPMEM memory prototype indicates a very high speed-up (two orders of magnitude) compared with state-of-the-art software (specifically GATK).

7.5. Benchmarks and Reviews

7.5.1. Evaluation of error correction tools for long Reads

Participants: Lolita Lecompte, Camille Marchet, Pierre Peterlongo.

Long read technologies, 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.

ELECTOR is freely available at https://github.com/kamimrcht/ELECTOR. It as been presented during the Jobim2018 conference [32]

7.5.2. Computational pan-genomics: status, promises and challenges

Participant: Pierre Peterlongo.

We took part to the redaction of the review paper that proposes a state of the art of the pan-genomics current status, methods and future orientations. This paper has been published in *Briefings in Bioinformatics* [18].

7.6. Bioinformatics Analysis

7.6.1. Metagenomic analysis of pea aphid symbiotic communities

Participants: Cervin Guyomar, Fabrice Legeai, Claire Lemaitre.

We worked on a methodological framework adapted to the study of genomic diversity and evolutionary dynamics of the pea aphid symbiotic community from an extensive set of metagenomics datasets. The framework is based on mapping to reference genomes and whole genome SNP-calling. We explored the genotypic diversity associated to the different symbionts of the pea aphid at several scales : across host biotypes, amongst individuals of the same biotype, and within individual aphids. Thorough phylogenomic analyses highlighted that the evolutionary dynamics of symbiotic associations strongly varied depending on the symbiont, reflecting different evolutionary histories and possible constraints [14].

7.6.2. Analysis of pea aphid genomic polymorphism

Participants: Fabrice Legeai, Claire Lemaitre.

We participated in the analyses of a large re-sequencing dataset of pea aphid individuals and populations. We performed the data cleaning, mapping to the reference genome and variant calling steps. The resulting polymorphism data shed light on two novel findings regarding the pea aphid genome evolution.

First, we showed that relaxed selection is likely to be the greatest contributor to the faster evolution of the X chromosome compared to autosomes [15]. Secondly, we looked for genomic bases of adaptation to novel environments, and identified 392 genomic hotspot regions of differentiation spanning 47.3 Mb and 2,484 genes. Interestingly, these hotspots were significantly enriched for candidate gene categories that are related to host–plant selection and use. These genes represent promising candidates for the genetic basis of host–plant specialization and ecological isolation in the pea aphid complex [21].

7.6.3. A de novo approach to disentangle partner identity and function in holobiont systems

Participants: Camille Marchet, Pierre Peterlongo.

Study of meta-transcriptomic datasets involving non-model organisms represents bioinformatic challenges that affect the study of holobiont meta-transcriptomes. Hence, we proposed an innovative bioinformatic approach and tested it on marine models as a proof of concept.

We considered three holobiont models, of which two transcriptomes were previously published and a yet unpublished transcriptome, to analyze and sort their raw reads using Short Read Connector (see section 7.1.1). Before assembly, we thus defined four distinct categories for each holobiont meta-transcriptome: host reads, symbiont reads, shared reads, and unassigned reads. Afterwards, we observed that independent *de novo* assemblies for each category led to a diminution of the number of chimeras compared to classical assembly methods. Moreover, the separation of each partner's transcriptome offered the independent and comparative exploration of their functional diversity in the holobiont. Finally, our strategy allowed to propose new functional annotations for two well-studied holobionts (a Cnidaria-Dinophyta, a Porifera-Bacteria) and a first meta-transcriptome from a planktonic Radiolaria-Dinophyta system forming widespread symbiotic association for which our knowledge is considerably limited [19].

7.6.4. Whole genome detection of micro-satellites

Participant: Dominique Lavenier.

This study has been done in cooperation with the federal university of de São João del-Rei, Brazil. The objective was to locate tens of thousands of micro-satellite loci for an endangered piracema (i.e. migratory) South American fish, *Brycon orbignyanus*. Together with the Brazil group we designed a specific pipeline that first assembles short paired-end reads into contigs and then performs micro-satellite oriented scaffolding processing [23].

7.6.5. Analysis of the genes and genomes involved in plant and insects interactions

Participant: Fabrice Legeai.

These study has been done in cooperation with various laboratories. In particular, we characterized the effectors (secreted proteins suppressing plant defense) of the pea aphid fed on different plants, by firstly identifying these genes in the pea aphid genome, then studying and comparing their expression between different conditions, and then finally by observing their evolution among a broad set of phytophagous insects [10]. We also identified microRNAs from smallRNA datasets from *Spodoptera frugiperda* strains fed on different host-plants [20]. Finally, we predicted the transposable elements in the genome of *Cephus cinctus*, an important insect pest [22].

7.6.6. Analysis of the expression and identification of the targets of mir202 during the medaka oogenesis

Participant: Fabrice Legeai.

This study has been done in cooperation with the INRA LPGP laboratory (Rennes). Its goal was to identify the role of small non-coding RNAs in the regulation of the reproduction of the fish model *Oryzias latipes* (medaka). We predicted the putative targets of the microRNA miR202, already observed as being specifically expressed in gonads. In the second part of the work, we identified important genes and functions targeted by miR202 and differentially expressed in the gonads when the microRNA was artificially repressed [13].

8. Bilateral Contracts and Grants with Industry

8.1. Bilateral Contracts with Industry

8.1.1. Processing in memory

Participants: Charles Deltel, Dominique Lavenier.

The UPMEM company is currently developing new memory devices with embedded computing power (http://www.upmem.com/). GenScale investigates how bioinformatics algorithms can benefit from these new types of memory. In 2018 we parallelized the detection of short variants (see new results section).

8.1.2. Tank milk analysis

Participants: Dominique Lavenier, Jacques Nicolas.

The Seenergi company has developed a biotechnology protocol to detect cow mastitis directly by analyzing 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, Sebastien Letort, 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, in collaboration with Biogemma's bioinformatics team, 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, Gregoire 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 Nanopore Minion, for rapid and robust pan-genome dicrimination of bacterial strains and their phenotypes. It has started at the end of this year with the recruitment (délégation Inria) of E. Roux, a biochemist from Lorraine University and G. Siekaniec (INRA -Inria collaboration, INRA grant), a new PhD student. We will study pan-genomic representations of multiple genomes and the production of characteristic signatures of each genome in this context.

9.2. National Initiatives

9.2.1. ANR

9.2.1.1. Project HydroGen: Metagenomic 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. 2018)
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 available from the Tara Oceans expedition.

9.2.1.2. Project SpeCrep: speciation processes in butterflies

Participants: Dominique Lavenier, Jeremy Gauthier, 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 – Dec. 2018)
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: 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 controlling adaptive mimicry in a polymorphic, ubiquitous 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.

9.2.2. PIA: Programme Investissement d'Avenir

9.2.2.1. RAPSODYN: Optimization of the rapeseed oil content under low nitrogen
Participants: Dominique Lavenier, Claire Lemaitre, Sebastien Letort, Pierre Peterlongo, Guendal 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 and 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, Celine Le Beguec.

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

The IPL Neuromarkers aims to design imaging bio-markers of neuro-degenerative diseases for clinical trials and study of their genetic associations. In this project, GenScale bring its expertise in the genomics field. More precisely, given a case-control population, a first step is to locate small genetic variations (SNPs, small indels) from their genomes. Then, having 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

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.

9.4. International Initiatives

9.4.1. Inria Associate Teams

9.4.1.1. HipcoGen

Title: High-Performance Combinatorial Optimization for Computational Genomics

International Partner (Institution - Laboratory - Researcher):

LANL (United States)

Information Science department

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.

9.4.1.2. 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 2018.
- Visit of Bernardo Clavijo from the Earlham Institute, United Kingdom, February 2018.
- Visit of Nicole Van Dam from Institute of Ecology, Jena university, June 2018.

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, 2018.
- Visit of S. Francois at Los Alamos National Laboratory, USA, from March 23 to April 23th, 2018.

10. Dissemination

10.1. Promoting Scientific Activities

10.1.1. Scientific Events Selection

10.1.1.1. Member of the Conference Program Committees

- JOBIM 2018: French symposium of Bioinformatics [P. Peterlongo]
- 2018 IEEE International Conference on Bioinformatics and Biomedicine [D. Lavenier]
- 9th International Workshop on Biological Knowledge Discovery from Big Data [D. Lavenier]
- International Work-Conference on Bioinformatics and Biomedical Engineering [D. Lavenier]
- 11th International Conference on Bioinformatics and Computational Biology (BICOB-2019) [R. Andonov]

10.1.1.2. Reviewer

- RECOMB 2018 [P. Peterlongo]
- SEQBIO 2018 [P. Peterlongo]
- RCAM 2018 [P. Peterlongo]
- ECCB 2018 [P. Peterlongo]

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]
- Current Bioinformatics [D. Lavenier]
- IEEE/ACM Transactions on Computational Biology and Bioinformatics [D. Lavenier]
- Nature Scientific Reports [F. Legeai]

10.1.3. Invited Talks

- C. Marchet, *A highly scalable data structure for read similarity computation and its application to marine holobionts*, EEB group meeting, Brussels (Belgium), July 8th 2018
- C. Marchet, *CARNAC-LR: clustering genes expressed variants from long read RNA sequencing*, team TIBS seminar, Rouen (France), February 27 2018
- P. Peterlongo, *DiscoSnp++, de novo Variant predictions*, Tara Oceans Polar Circle Consortium, Paris (France), January 2018
- P. Peterlongo, *Extract information from short metagenomic reads*, Workshop GDR GE Rennes, 4e colloque Génomique environnementale, October 2018.
- D. Lavenier, *GenScale and its life science interactions*, bioinformatics workshop, IFREMER, June 2018

10.1.4. Leadership within the Scientific Community

- P. Peterlongo. member of the Scientific Advisory Board of the GDR BIM (National Research Group in Biology, Informatic and Mathematic)
- C. Lemaitre. Animator of the Sequence Algorithms axis (seqBIM GT) of the GDR BIM (National Research Group in Biology, Informatics and Mathematics)
- F. Legeai. Animator of the INRA Center for Computerized Information Treatment "BBRIC".

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]
- Chapter "Artificial Intelligence and Bioinformatics" to appear in 2019 in the 3-volume book "A Guided Tour of Artificial Intelligence Research", H. Prade, P. Marquis and O. Papini eds, Elsevier. [35]

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) [J. Nicolas]
- Member of the steering committee of the INRA BIPAA Platform (BioInformatics Platform for Agroecosystems 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 UMR IRISA [C. Lemaitre]
- AGOS first secretary [P. Peterlongo]
- Organisation of the weekly seminar "Symbiose" [P. Peterlongo]
- 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]
- Representative of non-permanent members in the Inria Rennes center commitee [S. Letort]

10.2. Teaching - Supervision - Juries

10.2.1. Teaching

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

Licence : W. Delage, J. Gautier, Biostatistic, 60 h, L2, Univ. Rennes 1, France.

Licence : W. Delage, Bioinformatics, 8 h , L2, Univ. Rennes 1, France.

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

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

Master : L. Lecompte, Python for ecologists, 24h, M1, Univ. Rennes 1, France.

Master : L. Lecompte, P. Peterlongo, Algorithms on Sequences, 52h, M2, Univ. Rennes 1, France.

Master : C. Guyomar, statistical learning, 32h, M1, Univ. Rennes, France.

Master : P. Peterlongo, Experimental Bioinformactics, 24h, M1, ENS Rennes, France.

Master : F. Legeai, RNA-Seq, Metagenomics and Variant discovery, 12h, M2, AgroCampusOuest, 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
Training : C. Lemaitre, P. Peterlongo, GATB Programming Day, 8h (April 18), Paris, France.

10.2.2. Supervision

PhD : C. Guyomar, Bioinformatic tools and applications for metagenomics of bacterial communities associated to insects, Univ Rennes, 07/12/2018, J.C. Simon, C. Mougel, C. Lemaitre. [8]
PhD : C. Marchet, Nouvelles méthodologies pour l'assemblage de données de séquençage polymorphes, Univ Rennes, 20/09/2018, P. Peterlongo. [9]
PhD in progress : S. François, Combinatorial Optimization Approaches for Bioinformatics, 01/10/2016, 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: un nouveau cadre pour l'exploration de données de séquençage du microbiote intestinal, 01/10/2018, M. Berland, N. Pons and P. Peterlongo.
PhD in progress: R. Kar, Assembly and annotation of heterozygous insect genomes, 15/08/2018, F. Legeai, D. Tagu and P. Peterlongo.
PhD in progress: G. Siekaniec, Caractérisation différentielle de souches d'espèces bactériennes, 01/10/2018, E. Roux and J. Nicolas.

10.2.3. Juries

- *Member of Habilitation thesis jury*. Thomas Bruls, Univ. Paris-Saclay [D. Lavenier]
- *Member of Ph-D thesis juries*. Yoann Seelheuthner, University Paris-Saclay [C. Lemaitre].
- *Referee of Ph-D thesis*. Antoine Recanati, ENS Paris [D. Lavenier], Nicolas Dierckxsens, Univ. Libre de Bruxelle [D. Lavenier], Chadi Saad, Univ. Lille [J. Nicolas], Florian Plaza Oñate, INRA MetaGenoPolis [P. Peterlongo]
- *Member of Ph-D thesis committee*. Chi Nguyen Lam, Univ. Brest [D. Lavenier], Benjamin Churcheward, Univ. Nantes [D. Lavenier], Guillaume Gautreau, CEA, Genoscope [P. Peterlongo], Victor Gaborit, Univ. Nantes [P. Peterlongo], Afaf Saaid, Univ. Paris Saclay, Ecole Polytechnique [P. Peterlongo], Magali Dancette, Univ. Claude Bernard Lyon 1 [P. Peterlongo]

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]

10.3.3. Creation of media or tools for science outreach

- Short Movie "Les mutations génétiques, des événements spontanés à la thérapie génique", presented at Sciences en Courts, a local contest of popularization short movies made by PhD students (http://sciences-en-courts.fr/) [L. Lecompte, W. Delage].

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^{o} 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] 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

[6] 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

[7] 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

Doctoral Dissertations and Habilitation Theses

[8] C. GUYOMAR. *Bioinformatic tools and applications for metagenomics of microbial communities associated to insects*, Université Rennes 1, December 2018, https://tel.archives-ouvertes.fr/tel-01955222

[9] 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

Articles in International Peer-Reviewed Journals

[10] H. BOULAIN, F. LEGEAI, E. GUY, S. MORLIERE, N. DOUGLAS, J. OH, M. MURUGAN, M. SMITH, J. JAQUIÉRY, J. PECCOUD, F. WHITE, J. CAROLAN, J.-C. SIMON, A. SUGIO. *Fast Evolution and Lineage-Specific Gene Family Expansions of Aphid Salivary Effectors Driven by Interactions with Host-Plants*, in "Genome Biology and Evolution", 2018, vol. 10, n° 6, pp. 1554-1572 [*DOI :* 10.1093/GBE/EVY097], https://hal.archives-ouvertes.fr/hal-01891942

[11] C. BENOIT-PILVEN, C. MARCHET, E. CHAUTARD, L. LIMA, M.-P. LAMBERT, G. SACOMOTO, A. REY, A. COLOGNE, S. TERRONE, L. DULAURIER, J.-B. CLAUDE, C. BOURGEOIS, D. AUBOEUF, V. LACROIX. *Complementarity of assembly-first and mapping-first approaches for alternative splicing annotation and differential analysis from RNAseq data*, in "Scientific Reports", December 2018, vol. 8, n° 1 [*DOI :* 10.1038/s41598-018-21770-7], https://hal.inria.fr/hal-01924204

[12] G. CHAPUIS, H. DJIDJEV, G. HAHN, G. RIZK. *Finding Maximum Cliques on the D-Wave Quantum Annealer*, in "Journal of Signal Processing Systems", May 2018 [*DOI :* 10.1007/s11265-018-1357-8], https://hal.archives-ouvertes.fr/hal-01920397

[13] S. GAY, J. BUGEON, A. BOUCHAREB, L. HENRY, C. DELAHAYE, F. LEGEAI, J. MONTFORT, A. LE CAM, A. SIEGEL, J. BOBE, V. THERMES. *MiR-202 controls female fecundity by regulating medaka oogenesis*, in "PLoS Genetics", September 2018, vol. 14, n^{o} 9, pp. 1-26 [*DOI :* 10.1371/JOURNAL.PGEN.1007593], https://hal.archives-ouvertes.fr/hal-01871468

[14] 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^{o} 1, pp. 1-21 [*DOI :* 10.1186/s40168-018-0562-9], https://hal.archives-ouvertes.fr/hal-01926402

[15] J. JAQUIÉRY, J. PECCOUD, T. OUISSE, F. LEGEAI, N. PRUNIER-LETERME, A. GOUIN, P. NOUHAUD, J. A. BRISSON, R. BICKEL, S. PURANDARE, J. POULAIN, C. BATTAIL, C. LEMAITRE, L. MIEUZET, G. LE TRIONNAIRE, J.-C. SIMON, C. RISPE. *Disentangling the Causes for Faster-X Evolution in Aphids*, in "Genome Biology and Evolution", January 2018, vol. 10, n^{o} 2, pp. 507-520 [*DOI :* 10.1093/GBE/EVY015], https://hal.archives-ouvertes.fr/hal-01701165

[16] C. MARCHET, L. LECOMPTE, C. DA SILVA, C. CRUAUD, J.-M. AURY, J. NICOLAS, P. PETERLONGO. *De Novo Clustering of Long Reads by Gene from Transcriptomics Data*, in "Nucleic Acids Research", 2018, pp. 1-12 [*DOI :* 10.1093/NAR/GKY834], https://hal.archives-ouvertes.fr/hal-01643156

[17] C. MARCHET, L. LECOMPTE, A. LIMASSET, L. BITTNER, P. PETERLONGO. *A resource-frugal probabilistic dictionary and applications in bioinformatics*, in "Discrete Applied Mathematics", April 2018, pp. 1-11 [*DOI :* 10.1016/J.DAM.2018.03.035], https://hal.archives-ouvertes.fr/hal-01873312

[18] T. MARSCHALL, M. MARZ, T. ABEEL, L. DIJKSTRA, B. E. DUTILH, A. GHAFFAARI, P. KERSEY, W. P. KLOOSTERMAN, V. MAKINEN, A. M. NOVAK, B. PATEN, D. PORUBSKY, E. RIVALS, C. ALKAN, J. A. BAAIJENS, P. I. W. DE BAKKER, V. BOEVA, R. J. P. BONNAL, F. CHIAROMONTE, R. CHIKHI, F. D. CICCARELLI, R. CIJVAT, E. DATEMA, C. M. V. DUIJN, E. E. EICHLER, C. ERNST, E. ESKIN, E. GARRISON, M. EL-KEBIR, G. W. KLAU, J. O. KORBEL, E.-W. LAMEIJER, B. LANGMEAD, M. MARTIN, P. MEDVEDEV, J. C. MU, P. NEERINCX, K. OUWENS, P. PETERLONGO, N. PISANTI, S. RAHMANN, B. RAPHAEL, K. REINERT, D. DE RIDDER, J. DE RIDDER, M. SCHLESNER, O. SCHULZ-TRIEGLAFF, A. D. SANDERS, S. SHEIKHIZADEH, C. SHNEIDER, S. SMIT, D. VALENZUELA, J. WANG, L. WESSELS, Y. ZHANG, V. GURYEV, F. VANDIN, K. YE, A. SCHÖNHUTH. *Computational pan-genomics: status, promises and challenges*, in "Briefings in Bioinformatics", 2018, vol. 19, n^{o} 1, pp. 118-135 [*DOI :* 10.1093/BIB/BBW089], https://hal.inria.fr/hal-01390478

[19] A. MENG, C. MARCHET, E. CORRE, P. PETERLONGO, A. ALBERTI, C. DA SILVA, P. WINCKER, E. PELLETIER, I. PROBERT, J. DECELLE, S. LE CROM, F. NOT, L. BITTNER. *A de novo approach to disentangle partner identity and function in holobiont systems*, in "Microbiome", June 2018, pp. 1-35 [*DOI :* 10.1101/221424], https://hal.archives-ouvertes.fr/hal-01643153

[20] Y. MONÉ, S. NHIM, S. GIMENEZ, F. LEGEAI, I. SÉNINET, H. PARRINELLO, N. NEGRE, E. D'ALENÇON. *Characterization and expression profiling of microRNAs in response to plant feeding in two host-plant strains of the lepidopteran pest Spodoptera frugiperda*, in "BMC Genomics", December 2018, vol. 19, n^{o} 1, pp. 1-15 [*DOI :* 10.1186/s12864-018-5119-6], https://hal.inria.fr/hal-01926342

[21] P. NOUHAUD, M. GAUTIER, A. GOUIN, J. JAQUIÉRY, J. PECCOUD, F. LEGEAI, L. MIEUZET, C. SMADJA, C. LEMAITRE, R. VITALIS, J.-C. SIMON. *Identifying genomic hotspots of differentiation and candidate genes involved in the adaptive divergence of pea aphid host races*, in "Molecular Ecology", August 2018, vol. 27, n° 16, pp. 3287 - 3300 [*DOI :* 10.1111/MEC.14799], https://hal.archives-ouvertes.fr/hal-01892027

[22] R. ROBERTSON, R. WATERHOUSE, K. WALDEN, L. RUZZANTE, M. REIJNDERS, B. COATES, F. LEGEAI, J. GRESS, S. BIYIKLIOGLU, D. WEAVER, K. WANNER, H. H. BUDAK. *Genome sequence of the wheat stem sawfly, Cephus cinctus, representing an early-branching lineage of the Hymenoptera, illuminates evolution of hymenopteran chemoreceptors*, in "Genome Biology and Evolution", October 2018, vol. 10, n° 11, pp. 2997–3011 [*DOI :* 10.1093/GBE/EVY232], https://hal.inria.fr/hal-01926352

[23] G. YAZBECK, R. S. OLIVEIRA, J. M. RIBEIRO, R. GRACIANO, R. SANTOS, F. CARMO, D. LAVENIER. *A broad genomic panel of microsatellite loci from Brycon orbignyanus (Characiformes: Bryconidae) an endangered migratory Neotropical fish*, in "Scientific Reports", December 2018, vol. 8, n° 1, pp. 1-5 [*DOI :* 10.1038/S41598-018-26623-X], https://hal.archives-ouvertes.fr/hal-01919836

International Conferences with Proceedings

[24] F. COSTE, J. NICOLAS. *Learning local substitutable context-free languages from positive examples in polynomial time and data by reduction*, in "ICGI 2018 - 14th International Conference on Grammatical Inference", Wrocław, Poland, September 2018, vol. 93, pp. 155 - 168, https://hal.inria.fr/hal-01872266

[25] S. FRANCOIS, R. ANDONOV, H. DJIDJEV, M. TRAIKOV, N. YANEV. *Mixed Integer Linear Programming Approach for a Distance-Constrained Elementary Path Problem*, in "CTW 2018 - 16th Cologne-Twente Workshop on Graphs and Combinatorial Optimization", Paris, France, June 2018, pp. 1-4, https://hal.inria.fr/hal-01937008

[26] *Best Paper*
S. FRANCOIS, 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.

[27] A. LIMASSET, J.-F. FLOT, P. PETERLONGO. *Toward perfect reads: self-correction of short reads via mapping on de Bruijn graphs*, in "Recomb", Paris, France, April 2018, https://arxiv.org/abs/1711.03336 - RECOMB 2018 Submission, https://hal.inria.fr/hal-01644163

Conferences without Proceedings

[28] C. MARCHET, L. LECOMPTE, C. DA SILVA, C. CRUAUD, J.-M. AURY, J. NICOLAS, P. PETERLONGO. *CARNAC-LR : Clustering coefficient-based Acquisition of RNA Communities in Long Reads*, in "JOBIM 2018 - Journées Ouvertes Biologie, Informatique et Mathématiques", Marseille, France, July 2018, pp. 1-3, https://hal.archives-ouvertes.fr/hal-01930211

Research Reports

[29] M. BRIDOUX. *Séparation d'haplotypes à partir de données de séquençage de troisième génération*, Inria Rennes - Bretagne Atlantique, August 2018, https://hal.archives-ouvertes.fr/hal-01933561

Other Publications

[30] 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, G. V. MARKOV. *Inferring biochemical reactions and metabolite structures to cope with metabolic pathway drift*, December 2018, working paper or preprint, https://hal.inria.fr/hal-01943880

[31] C. GUYOMAR, W. DELAGE, F. LEGEAI, C. C. MOUGEL, J.-C. SIMON, C. LEMAITRE. *Reference guided genome assembly in metagenomic samples*, April 2018, 1 p. , RECOMB 2018 - 22nd International Conference on Research in Computational Molecular Biology, Poster, https://hal.archives-ouvertes.fr/hal-01934823

[32] L. LECOMPTE, C. MARCHET, P. MORISSE, A. LIMASSET, P. PETERLONGO, A. LEFEBVRE, T. LECROQ. *ELECTOR: EvaLuation of Error Correction Tools for lOng Reads*, July 2018, pp. 1-2, JOBIM 2018 - Journées Ouvertes Biologie, Informatique et Mathématiques, Poster, https://hal.archives-ouvertes.fr/hal-01929900

[33] C. MARCHET, L. LECOMPTE, C. DA SILVA, C. CRUAUD, J.-M. AURY, J. NICOLAS, P. PETERLONGO. *CARNAC-LR: De novo Clustering of Gene Expressed Variants in Transcriptomic Long Reads Data Sets*, April 2018, pp. 1-2, RECOMB-seq 2018 - Eighth RECOMB Satellite Workshop on Massively Parallel Sequencing, Poster, https://hal.archives-ouvertes.fr/hal-01929963

[34] C. MARCHET, L. LIMA. *Simulation of RNA sequencIng with Oxford Nanopore Technologies*, July 2018, pp. 1-2, JOBIM 2018 - Journées Ouvertes Biologie, Informatique et Mathématiques, Poster, https://hal.archives-ouvertes.fr/hal-01929917

[35] J. NICOLAS. *Artificial Intelligence and Bioinformatics*, July 2018, working paper or preprint, https://hal.inria.fr/hal-01850570

References in notes

[36] J. T. SIMPSON, K. WONG, S. D. JACKMAN, J. E. SCHEIN, S. J. M. JONES, I. BIROL. *ABySS: a parallel assembler for short read sequence data*, in "Genome Res", Jun 2009, vol. 19, n^{o} 6, pp. 1117–1123