IN PARTNERSHIP WITH:
CNRS

Université Rennes 1

École normale supérieure de Rennes

Activity Report 2017

Project-Team GENSCALE

Scalable, Optimized and Parallel Algorithms for Genomics

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

RESEARCH CENTER
Rennes - Bretagne-Atlantique

THEME
Computational Biology

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:

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:

B1.1.6. - Genomics
B1.1.9. - Bioinformatics
B2.2.3. - Cancer

1. Personnel

Research Scientists

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

Faculty Member

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

PhD Students

Gaetan Benoit [Inria]
Wesley Delage [Inria, from Oct 2017]
Sebastien Francois [Univ de Rennes I]
Cervin Guyomar [Univ de Rennes I]
Lolita Lecompte [Inria, from Sep 2017]
Antoine Limasset [Univ de Rennes I, until Aug 2017]
Camille Marchet [Univ de Rennes I]
Hoang Son Pham [Vietnam gov.]

Technical staff

Jennifer Del Giudice [Inria, until Mar 2017]
Patrick Durand [Inria, until Aug 2017]
Jeremy Gauthier [Inria, from Feb 2017]
Sebastien Letort [CNRS]
Stephane Picq [Inria, from Feb 2017 until Jun 2017]
Charles Deltel [Inria, 50%]

Interns

Wesley Delage [INRA, until Jun 2017]
Mael Kerbiriou [CNRS, from Feb 2017 until Jul 2017]
Lolita Lecompte [Inria, until Jun 2017]
Charlotte Mouden [Univ de Rennes I, from Apr 2017 until Sep 2017]

Administrative Assistant

Marie Le Roic [Univ de Rennes I]

Visiting Scientists

Hristo Djidjev [Los Alamos National Laboratory, from Jun 2017 until Jul 2017]
Metodi Traykov [Agence Erasmus+ France, from Mar 2017 until Apr 2017]

External Collaborators

Susete Alves Carvalho [INRA, from Mar 2017]
Fabrice Legeai [INRA]
Guillaume Rizk [AlgoRizk Company]

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 throughtput 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 NGS 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 NGS tools. 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) 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 [45]. 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 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.

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.

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.
- **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 and on the scaffolding step.
- **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.
- **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.

- **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.

3.3. Axis 3: Parallelism

This third axis is 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. The second is a parallel algorithmic approach: during the development step, the algorithms are constrained by parallel criteria. 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 N-traits 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.

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. CAMI

GenScale participated to the international CAMI challenge. CAMI stands for Critical Assessment of Metagenome Interpretation. It is a community-led initiative designed to tackle the problem of recovering the complex information encoded in metagenomes by aiming for an independent, comprehensive and bias-free evaluation of methods. We contributed in the "Assembly" section with the Minia pipeline. Results of this competition, presented in the "Nature Methods" journal [20], highlight the good behaviour of our tool compared to other competitors.

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++.

RELEASE FUNCTIONAL DESCRIPTION: speed up from x2 to x4 for kmer counting and graph construction phases (optimizations based on minimizers and improved Bloom filters). GATB's k-mer counter has been improved using techniques from KMC2, to achieve competitive running times compared to KMC2. ability to store arbitrary information associated to each kmer of the graph, enabled by a minimal perfect hash function (costs only 2.61 bits/kmer of memory) improved API with new possibilities (banks and kmers management) many new snippets showing how to use the library.

- Participants: Charles Deltel, Claire Lemaitre, Dominique Lavenier, Guillaume Rizk, Patrick Durand and Pierre Peterlongo
- Contact: Dominique Lavenier
- URL: http://gatb.inria.fr/

6.2. 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.

- Contact: Pierre Peterlongo
- URL: https://github.com/GATB/DiscoSnp

6.3. GWASDM

Genome Wide Association Study using Data Mining strategy
KEYWORDS: GWAS - Data mining
FUNCTIONAL DESCRIPTION: From two cohorts of genotyped individuals (case and control), the GWASDM software performs a Genome Wide Association Study based on data mining techniques and generates several patterns of SNPs that correlate with a given phenotype. The algorithm implemented in GWASDM directly uses relative risk measures such as risk ratio, odds ratio and absolute risk reduction combined with confidence intervals as anti-monotonic properties to efficiently prune the search space. The algorithm discovers a complete set of discriminating patterns with regard to given thresholds or applies heuristic strategies to extract the largest statistically significant discriminating patterns in a given dataset.

- Contact: Dominique Lavenier

6.4. bcool

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

- Partner: Université libre de Bruxelles
- Contact: Pierre Peterlongo
- URL: http://github.com/Malfoy/BCOOL

6.5. 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 describes a 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.

- Contact: Camille Marchet

7. New Results

7.1. Data Structure

7.1.1. Minimal perfect hash function

Participants: Antoine Limasset, Guillaume Rizk, Pierre Peterlongo.

Minimal perfect hash functions are fundamental objects used in many applications. Existing algorithms and implementations that build such functions have in practice some upper bounds on the number of input elements they can handle, due to high construction time and/or memory usage. We propose a simple algorithm having very competitive construction times, memory usage and query times compared to state of the art techniques [27]. We provide a parallel implementation called BBHash. It is capable of creating a minimal perfect hash function of 10^{10} elements in less than 1 hour and 4 GB of memory. To the best of our knowledge, this library is also the first that has been successfully tested on 10^{12} input elements. Source code: https://github.com/rizkg/BBHash

7.1.2. Quasi-dictionary

Participants: Camille Marchet, Antoine Limasset, 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. In this context, we proposed a probabilistic data structure based on a minimal perfect hash function for indexing large sets of keys. This structure out-competes 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 provided two applications based on similarity computation between collections of sequences, and for which this calculation is an expensive but required operation. In particular, we showed a practical case in which other bioinformatics tools failed to scale up the tested data set or provide lower recall quality results [43].

7.2. Algorithms & Methods

7.2.1. Short Read Correction

Participants: Antoine Limasset, 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 $k$mer-spectrum techniques, while being scalable, making it possible to apply it to human-size genomic datasets and beyond [41].

7.2.2. Long transcriptomic read clustering

Participants: Camille Marchet, Pierre Peterlongo.

This contribution tackles the problem of clustering RNA reads in clusters representing all variants of each gene, in a *de novo* way i.e. without any reference sequences. Such problem is not new as is, but the latest, Third Generation Sequencing (TGS) data redefine it. Reads can now span full-length transcripts but at the price of very high error rates, mostly insertions and deletions. This makes difficult or impossible to use tools designed for previous sequencing data. Still, the property to obtain whole RNA molecules through reads is very promising to better describe a transcriptome. In this work, we targeted the need to extract relevant information from a TGS transcriptome, even when no reference is available. In collaboration with Jacques Nicolas from the Inria/IRISA Dyliss team, we propose a novel algorithm in the community detection framework, based on the clustering coefficient. In addition we propose an implementation of this algorithm in the tool CARNAC-LR and a pipeline for the processing of transcriptome data. We validated our tool on real data from mouse and showed that it could be accurate and precise even for lowly expressed genes. We showed that our approach can be complementary to a mapping in the case a reference exists, and that a straightforward use of CARNAC-LR enables to quickly assess the genes'e expression levels [42].

7.2.3. Statistically Significant Discriminative Patterns Search

Participants: Hoang Son Pham, Dominique Lavenier.

Identifying multiple SNPs combinations associated with diseases such as cancers or diabetes is a central goal of human genetics. Recently, discriminative pattern mining algorithms have been investigated to tackle genome-wide association studies (GWAS). We designed an algorithm, called SSDPS, to discover groups of items which have significant difference of frequency in case-control datasets. The algorithm directly uses relative risk measures such as risk ratio, odds ratio and absolute risk reduction combined with confidence intervals as anti-monotonic properties to efficiently prune the search space. The algorithm discovers a complete set of discriminative patterns with regard to given thresholds or applies heuristic strategies to extract the largest statistically significant discriminative patterns in a given dataset. Experimental results on both synthetic datasets and three real variant datasets (Age-Related Macular Degeneration, Breast Cancer and Type 2 Diabetes) demonstrate that the SSDPS algorithm effectively detects multiple SNPs combinations in an acceptable execution time.

7.2.4. Reference free SNP detection in RAD-seq data

Participants: Jeremy Gauthier, Claire Lemaitre, Pierre Peterlongo.

We developed an original method for reference-free variant calling from Restriction site associated DNA Sequencing (RAD-Seq) data. RAD-seq is a technique widely employed in the evolutionary biology field. Based on the variant caller DiscoSnp, DiscoSnp-RAD explores the De Bruijn Graph built from all the read datasets to detect SNP and Indels. Tested on simulated and real datasets, DiscoSnp-RAD identifies thousands of variants suitable for different population genomics analyses. Furthermore, DiscoSnp-RAD stands out from other tools due to his completely different principle, making it significantly faster, in particular on large datasets [39].

7.2.5. Global Optimization for Scaffolding and Completing Genome Assemblies

Participants: Sebastien Francois, Rumen Andonov, Dominique Lavenier.

We developed a method for solving genome scaffolding as a problem of finding the longest simple path in a graph defined by the contigs that satisfies a maximal number of additional constraints encoding the insert-size information [26]. Then we solved the resulting mixed integer linear program to optimality using the Gurobi solver. We tested our algorithm on a benchmark of chloroplast genomes and showed that it outperforms other widely-used assembly solvers by the accuracy of the results.

7.2.6. Identification and characterization of long non-coding RNA

Participant: Fabrice Legeai.

We participated in the development and validation of the tool FeelNC (collaboration with IGDR group). This is a tool allowing the identification of long non coding RNA (lncRNA) from RNASeq reads with or without a reference genome. Contrary to other tools, it does not depend on the comparison with protein databanks, which usually require lots of computations, but used a machine learning approach based on a Random Forest model trained with general features such as multi k-mer frequencies and relaxed open reading frames. We delivered a module that allows to characterize the relationships of each long non coding RNA with the other genes in its genomics close environment, giving insights about the putative impact of the lncRNAs to the regulation of these genes [23], [24].

7.2.7. Characterizing repeat-associated subgraphs in de Bruijn graphs

Participant: Camille Marchet.

The main problem in genome assembly, namely repeats, is also present in transcriptomic data. They are dealt with using various heuristics in the de Bruijn Graph framework (dBG). In this work, we introduce a formal model for representing high copy-number and low-divergence repeats in RNA-seq data in dBG and infer the definition of repeat-associated subgraphs. We show that the problem of identifying such subgraphs in a dBG is NP-complete. Then we place ourselves in the case of local assembly of alternative splicing and show that such subgraphs can be avoided implicitly. Thus, more alternative splicing events can be enumerated than with previous approaches. Finally we show that this exploration of DBG explorations can improve de novo transcriptome evaluation methods [16].

7.3. Parallelism

7.3.1. Variant detection using processing-in-memory technology

Participants: Charles Deltel, Dominique Lavenier.

The concept of Processing-In-Memory aims to dispatch the computer power near the data. Together with the UPMEM company (http://www.upmem.com/), which is currently developing a DRAM memory enhanced with computing units, we investigate the parallelization of the detection of mutations on the human genome. Traditionnaly, 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, allowing the variant calling step to be done simultaneously on the host processor. The implementation is currently on going. First performance evaluations indicate speed-up of one or two order of magnitude compared to purely software implementation.

7.4. Bioinformatics Analysis

7.4.1. Study of marine plankton holobionts

Participants: Camille Marchet, Pierre Peterlongo.

We derived from the quasi-dictionary (described in previous section) a tool called Short Read Connector (SRC), able to find pairs of similar reads intra or inter read sets. We used SRC in meta-transcriptomics context to identify the actors of a symbiosis and help the assembly [44], [31]. The framework is the study of marina holobionts (host and its community of symbionts) for which few is known about the actors. In order to retrieve the functions that characterize such holobionts, RNA-seq reads from the sequencing of the whole holobiont are assembled *de novo*. Such assembly is prone to produce chimeras. Thus SRC is used to index sequences (reads, EST, assembled genes...) known to be close to the host and symbionts of the holobiont. Then, thanks to SRC's ability to find similarity between sequences even at a large scale, by querying reads of the holobiont we identify those similar to the host or symbionts. We report four categories: host, symbiont, shared and unassigned that can be assembled in a parallel way. As a first step we validate the SRC+assembly approach by comparing our result to literature with two known holobionts with eukaryote hosts (*Orbicella faveolata, Xestospongia muta*). We show that our approach can compare to previous results. In a second step we lean on a protist (Collodaria) holobiont for which the actors are poorly known. No assembled sequences exist in the literature so we compare the pipeline SRC+assembly to a sole assembly pipeline. Our main achievement is to highly reduce (up to $\sim$40%) the number of chimeras in the assembly compared to the sole assembly pipeline.

7.4.2. Pea aphid metagenomics

Participants: Cervin Guyomar, Fabrice Legeai, Claire Lemaitre.

We worked on a 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, or within individual aphids. Thorough phylogenomic analyses highlighted that the evolutionary dynamics of symbiotic associations strongly varied depending on the symbiont, reflecting different histories and possible constraints [40], [30].

7.4.3. Assembly and comparison of two genomes of highly polyphagous lepidopteran pests

Participants: Fabrice Legeai, Claire Lemaitre.

In this study, two genomes of an agronomical important lepidopteran pest, the noctuid moth *Spodoptera frugiperda*, were sequenced and compared, giving significant insights to the mechanisms involved in host-plant adaptation and speciation of this organism. In particular, we described the large expansion of gustatory receptors and detoxification genes among this polyphagous pest compared to other specialist Lepidoptera, and emphasizes the role of these 2 gene families in the evolution of one of the world's worst agricultural pests. We also provided the genome assemblies, gene annotations and whole genome alignments of both strains, and the comparison of both to a reference moth genome (*Bombyx mori*). For these purposes, several original methods were developed i) to correct genome assembly errors due to the high level of heterozygosity and ii) to extract structural variant calls from whole genome alignments [15].

7.4.4. Benchmark of de novo read dataset compression tools

Participants: Gaetan Benoit, Dominique Lavenier, Claire Lemaitre.

In this book chapter, we review the different approaches and their tools developed so far to compress sequencing data files. We detail the algorithms for each of the three main types of data contained in such files for each read : the header, the DNA and the quality sequences. We also provide a thorough benchmark of the numerous available tools on various sequencing datasets, evaluating the compression ratio as well as the running time and memory usage performances [33].

7.4.5. Genomics of the agro-ecosystems pests

Participants: Fabrice Legeai, Claire Lemaitre.

Within a large international network of biologists, GenScale has contributed to various projects for identifying important components such as protein coding or non coding genes involved in the adaptation of major agricultural pests to their environment. We provided or participated to the assembly and the annotation of 4 new aphids [17], [22], and 5 parasitic wasps. Following specific agreement or policy, these new genomes and annotations are available for a restricted consortium or a large community through the BioInformatics platform for Agro-ecosystems Arthropods (http://bipaa.genouest.org/is). Moreover our engagement in the agronomical pest genomics led to our contribution to other projects such as epigenetics and chromatin structure analysis [18], or the analysis of population genetics data for identifying hotspots of selection in the nematode *Globodera pallida* genome [14].

7.4.6. 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 the multi-mapping results. 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 [37].

7.4.7. Microbial communities interaction between plant and their bioagressors

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

GenScale actively collaborates with the INRA group ‘plant-microbial communities interactions’ (IGEPP, Rennes) that analyze the interaction between plant, their associated microbial communities and different bioagressors. The ambition of the project is to understand the link between the taxonomic biodiversity of the microbiota and their functional diversity in relation with plant physiology and plant-bioagressors interactions. For this last point, an integrated metatranscritomic approach is developped. Beside wet lab and sequence productions, bioinformatics tools are needed and meta-transcriptomic pipelines analysis arecurrently developped based on the GenScale expertise.

7.5. Challenges

7.5.1. Participation to CAMI: de-novo metagenomics assembly competition

Participants: Charles Deltel, Dominique Lavenier, Claire Lemaitre, Pierre Peterlongo.

In metagenome analysis, computational methods for assembly, taxonomic profiling and binning are key components facilitating downstream biological data interpretation. However, a lack of consensus about benchmarking datasets and evaluation metrics complicates proper performance assessment. In this context, we participated to CAMI (Critical Assessment of Metagenome Interpretation), specifically on the assembly section with the Minia pipeline. The CAMI challenge aimed to benchmark programs on datasets of unprecedented complexity and realism. Benchmark metagenomes were generated from 700 newly sequenced microorganisms and 600 novel viruses and plasmids, including genomes with varying degrees of relatedness to each other and to publicly available ones and representing common experimental setups. Across all datasets, our assembly programs performed well for species represented by individual genomes, while performance was substantially affected by the presence of related strains [20].

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 (see section New Results).

8.2. Bilateral Grants with Industry

8.2.1. Enancio Start-Up

Participants: Jennifer Del Giudice, Stephane Picq, Guillaume Rizk.

After 2 years of development the EnginesOn project has led to the creation of Enancio in August 2017 (http://www.enancio.fr). Enancio main focus is to give the biologist all the resources needed to decipher the information held on a biological molecule such as DNA, without worrying about the informatics behind it. The start-up provides a software platform available through the net with analysis workflows that have been conceived and validated by the field experts, solutions to handle massive data, and health data certified computational infrastructure. Simplification, optimization and faster execution of analyses workflows are the main focuses of the company. Enancio workflows uses the GATB-core library developed by GenScale.

8.2.2. Rapsodyn project

Participants: Dominique Lavenier, Claire Lemaitre, Sebastien Letort, Pierre Peterlongo.

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. Rennes Hospital, Hematology service

Participants: Dominique Lavenier, Patrick Durand.

The collaboration with the Hematology service of the Rennes hospital aims to set up advanced bioinformatics pipelines for cancer diagnosis. More precisely, we evaluated a new method of predictions of small cancer-related mutations (such as SNPs and small insertions/deletions) from raw DNA sequencing data.

9.1.2. Partnership with INRA in Rennes

Participants: Susete Alves Carvalho, Cervin Guyomar, Dominique Lavenier, Fabrice Legeai, Claire Lemaitre, Sebastien Letort, Pierre Peterlongo.

The GenScale team has a strong and long term collaboration with biologists of INRA in Rennes: IGEPP and PEGASE units. This partnership concerns both service and research activities and is acted by the hosting of two INRA engineer (F. Legeai, S. Alves Carvalho) and one PhD student (C. Guyomar).

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, Guillaume Rizk, Gaetan Benoit.

Coordinator: P. Peterlongo (Inria/Irisa, GenScale, Rennes)
Duration: 42 months (Nov. 2014 – Apr. 2018)
Partners: CEA (GenosScope, 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 Systematique et d'Evolution de la Biodiversite, 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.

9.2.2. 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.

Coordinator: N. Nesi (Inra, IGEPP, Rennes)
Duration:7.5 year (2012-2019)
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)

9.2.2.2. Institut Français de Bioinformatique: Plant node

Participant: Fabrice Legeai.

Coordinator: Hadi Quesneville (INRA, Versailles)

The aim of the Institut Francais de Bioinformatique (IFB) offers resources for a large community of French biologist. With INRA and CIRAD, we were part of the plant node of IFB, and focused on delivering efficient tools for sharing agronomical data, such as Askomics.

9.2.3. Programs from research institutions

9.2.3.1. Inria ADT DiagCancer

Participants: Dominique Lavenier, Patrick Durand.

Since October 1st, 2016, Genscale started a one-year Inria ADT called DiagCancer. It aims at: (1) including the DiscoSnp++ tool within the current data production pipeline at Pontchaillou Hospital (Rennes), (2) providing a new prediction tool applied to the calling of cancer related mutations from DNA sequencing data and (3) creating new analysis tools to facilitate the interpretation of results by end-users (biologists, doctors). The project is done in close collaboration with Haematology Service, CHU Pontchaillou, Rennes.

9.2.3.2. CNRS Mastodons program: C3G

Participants: Dominique Lavenier, Pierre Peterlongo, Claire Lemaitre, Camille Marchet, Lolita Lecompte.

High-throughput sequencing applications now cover all life sciences: from medicine to agronomy. The 3rd generation sequencing produces very long reads, but the reads are extremely noisy, which has a strong impact on the quality of bioinformatics analyses. The challenge of the C3G project is to bring this type of data to a high level of quality through the development of new correction strategies.

9.2.3.3. Inria Project Lab: Neuromarkers

Participants: Dominique Lavenier, Pierre Peterlongo, Claire Lemaitre.

The IPL Neuromarkers aims to design imaging biomarkers of neurodegenerative diseases for clinical trials and study of their genetic associations. In this project, GenScale bring its expertise in the genomic field.

9.3. International Initiatives

9.3.1. Inria Associate Team: HipcoGen

- Title: High-Performance Combinatorial Optimization for Computational Genomics
- International Partner (Institution - Laboratory - Researcher):
  - Los Alamos National Laboratory (LANL)-NM, United States, CCS-3, Hristo Djidjev
- Start year: 2017
- Teams' web site: 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.

9.3.2. Informal International Partners

- Free University of Brussels, Belgium: Genome assembly [P. Perterlongo, A. Limasset]

9.4. International Research Visitors

9.4.1. Visits of International Scientists

- Visit of Hristo Djidjev from Los Alamos National Laboratory, June 5 to July 4, 2017

9.4.2. Visits to International Teams

- Visit of R. Andonov at LANL from May 4th to May 30th. Work on Task 2 from HipcoGen project.
- Visit of S. Francois at LANL from May 4th to May 30th and from August 2 to August 23. Work on Task 2 from HipcoGen project.
- Visit of Pierre Peterlongo at LANL, May 2017 (one week). Talk to SFAF conference: "Assembly of heterozygous genomes".

10. Dissemination

10.1. Promoting Scientific Activities

10.1.1. Scientific Events Organisation

10.1.1.1. Member of the Organizing Committees

- RCAM 2017: Workshop “Recent Computational Advances in Metagenomics" [P. Peterlongo]

10.1.2. Scientific Events Selection

10.1.2.1. Member of the Conference Program Committees

- JOBIM 2017: French symposium of Bioinformatics [C. Lemaitre, P. Peterlongo]
- BIBM 2017: IEEE International Conference on Bioinformatics and Biomedicine [D. Lavenier]
- BIOKDD 2017: Workshop on Biological Knowledge Discovery and Data Mining [D. Lavenier]
- IWBBIO 2017: International Work-Conference on Bioinformatics and Biomedical Engineering [D. Lavenier]

10.1.2.2. Reviewer

- RECOMB 2017 [R. Andonov, C. Lemaitre, P. Peterlongo]
- ECCB 2017 [P. Peterlongo]

10.1.3. Journal

10.1.3.1. Reviewer - Reviewing Activities

- Bioinformatics [D. Lavenier]
- BMC Bioinformtics [D. Lavenier, C. Marchet]
- Briefing in Bioinformatics [D. Lavenier]
- Plos One [C. Lemaitre]
- Transactions on Computational Biology and Bioinformatics [C. Lemaitre]
- Genomics [F. Legeai]

10.1.4. Invited Talks

- D. Lavenier, *Dealing with mass of genomic data. From optimized data structures to advanced memory architectures*, BIATA 2017, Dioinformatics: from algorithms to application, St, Petersburg, Russia, Aug. 2017
- D. Lavenier, *DNA Mapping using Processing-in-Memory Architecture*, Cristal, Univ. Lille, Oct 2017
- C. Lemaitre, *Looking for genomic variants in the De Bruijn Graph.*, Institute for Advanced Biosciences, University of Grenoble Alpes, Grenoble, France, Dec. 2017
- C. Marchet, *De novo Clustering of Gene Expressed Variants in Transcriptomic Long Reads Data Sets*, at From RNA-Seq data to bioinformatics analysis using Nanopore sequencers, Evry, Dec 2017.

10.1.5. Leadership within the Scientific Community

- P. Peterlongo. Animator of the metagenomic scientific axes of the GDR BIM (National Researc Group in Biology, Informatic and Mathematic)
- P. Peterlongo. Member of the SFBI board.

10.1.6. Scientific Expertise

- Expert for the MEI (International Expertise Mission), French Research Ministry [D. Lavenier]
- Member of the Scientific Council of BioGenOuest [D. Lavenier]
- Member of the Scientific Council of the Computational Biology Institute of Montpellier [D. Lavenier]

10.1.7. Research Administration

- Member of the CoNRS, section 06, [D. Lavenier]
- Member of the CoNRS, section 51, [D. Lavenier]
- 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 Platform of BioGenOuest) [D. Lavenier]
- 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 : C. Lemaitre, Statistics for biology, 30h, L3, Univ. Rennes 1, France.

Licence : R. Andonov, S. Francois, Graph Algorithms, 80h, L3, Univ. Rennes 1, France.

Licence : L. Lecompte, Systems, 16h, L3, Univ. Rennes 1, France.

Licence : S. Francois, Programming, 22h, L3 Miage, Univ. Rennes 1, France.

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

Master : L. Lecompte, Python for ecologists, 21h, M1, Univ. Rennes 1, France.

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

Master : L. Lecompte, C. Lemaitre, P. Peterlongo, Algorithms on Sequences for Bioinformatics, 50h, M1, Univ. Rennes 1, France.

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

Master : C. Guyomar, Statistical learnings, 30h, M2, Univ. 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.

Training : P. Durand, G. Rizk, GATB Programming Day, 16h (April 26-27), Grenada, Spain.

Training : P. Durand, G. Rizk, GATB Programming Day, 8h (June 6), Montpellier, France.

Training : G. Rizk, GATB Programming Day, 8h (October 24), Rennes, France.

10.2.2. Supervision

PhD : G. Benoit, Large scale de novo comparative metagenomics, Univ Rennes, 29/11/2017, D. Lavenier and C. Lemaitre. [11]

PhD : A. Limasset, Nouvelles approches pour l'exploitation des données de séquençage haut débit, Univ Rennes, 12/07/2017, D. Lavenier and P. Peterlongo. [12]

PhD : P. Hoan Son, Novel Pattern Mining Techniques for Genome-wide Association Studies, 22/12/2017, D. Lavenier and A. Termier.

PhD in progress : C. Guyomar, Bioinformatic tools and applications for metagenomics of bacterial communities associated to insects, 01/10/2015, J.C. Simon, C. Mougel, C. Lemaitre and F. Legeai.

PhD in progress : C. Marchet, Nouvelles méthodologies pour l'assemblage de données de séquençage polymorphes, 01/10/2015, P. Peterlongo.

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, Assemblage de novo local pour la détection de variations complexes pour le diagnostic des maladies rares, 01/10/2017, J. Thévenon and C. Lemaitre.

10.2.3. Juries

- *Member of Ph-D thesis juries*. Arnaud Meng, University Pierre et Marie Curie [C. Lemaitre], Damien Courtine, University of Brest [D. Lavenier].
- *Referee of Ph-D thesis*. Pierre Pericard, University of Lille [D. Lavenier], Louise-Amelie Schmitt, Bordeaux University [P. Peterlongo], Kamil Salikov, University Paris Est [P. Peterlongo]

- *Member of Ph-D thesis comitees*. L. Ishi Soares de Lima, University of Lyon [C. Lemaitre], Cervin Guyomar, University of Rennes [P. Peterlongo], Pierre Marijon, University of Lille [P. Peterlongo], Pierre Charrier, Oniris Nantes [P. Peterlongo], Victor Gaborit, Inserm Nantes [P. Peterlongo], Guillaume Gautreau, University Paris Saclay [P. Peterlongo].
- *President of a jury for the recuitment of a INRA bioinformatics engineer* [Fabrice Legeai].

10.3. Popularization

- Participation to operation "A la découverte de la recherche" in high schools [P. Peterlongo]
- Intervention for "Imagine For Margo" [P. Peterlongo]
- Bioinformatics introduction to secondary school pupils [F. Legeai]

11. Bibliography

Major publications by the team in recent years

[1] R. ANDONOV, N. MALOD-DOGNIN, N. YANEV. *Maximum Contact Map Overlap Revisited*, in "Journal of Computational Biology", January 2011, vol. 18, n^o 1, pp. 1-15 [*DOI :* 10.1089/CMB.2009.0196], http://hal.inria.fr/inria-00536624/en

[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^o 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] N. MAILLET, C. LEMAITRE, R. CHIKHI, D. LAVENIER, P. PETERLONGO. *Compareads: comparing huge metagenomic experiments*, in "RECOMB Comparative Genomics 2012", Niterói, Brazil, October 2012, https://hal.inria.fr/hal-00720951

[6] N. MALOD-DOGNIN, R. ANDONOV, N. YANEV. *Maximum Cliques in Protein Structure Comparison*, in "SEA 2010 9th International Symposium on Experimental Algorithms", Naples, Italy, P. FESTA (editor), Springer, May 2010, vol. 6049, pp. 106-117 [*DOI :* 10.1007/978-3-642-13193-6_10], https://hal.inria.fr/inria-00536700

[7] V. H. NGUYEN, D. LAVENIER. *PLAST: parallel local alignment search tool for database comparison*, in "Bmc Bioinformatics", October 2009, vol. 10, 24 p. , http://hal.inria.fr/inria-00425301

[8] 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^o 24, pp. 3451 - 3457 [*DOI :* 10.1093/BIOINFORMATICS/BTU545], https://hal.inria.fr/hal-01081089

[9] G. RIZK, D. LAVENIER. *GASSST: Global Alignment Short Sequence Search Tool*, in "Bioinformatics", August 2010, vol. 26, n^{o} 20, pp. 2534-2540, http://hal.archives-ouvertes.fr/hal-00531499

[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

Doctoral Dissertations and Habilitation Theses

[11] G. BENOIT. *Large-scale de novo comparative metagenomics*, Université Rennes1, November 2017, https://hal.inria.fr/tel-01659395

[12] A. LIMASSET. *Novel approaches for the exploitation of high throughput sequencing data*, Université Rennes 1, July 2017, https://hal.archives-ouvertes.fr/tel-01566938

[13] H. S. PHAM. *Novel Pattern Mining Techniques for Genome-wide Association Studies*, IRISA, equipe GENSCALE, December 2017, https://hal.inria.fr/tel-01672442

Articles in International Peer-Reviewed Journals

[14] D. EOCHE-BOSY, M. GAUTIER, M. ESQUIBET, F. LEGEAI, A. BRETAUDEAU, O. BOUCHEZ, S. FOURNET, E. GRENIER, J. MONTARRY. *Genome scans on experimentally evolved populations reveal candidate regions for adaptation to plant resistance in the potato cyst nematode Globodera pallida*, in "Molecular Ecology", July 2017, vol. 26, n^{o} 18, pp. 4700–4711 [*DOI :* 10.1111/MEC.14240], https://hal.inria.fr/hal-01605681

[15] A. GOUIN, A. BRETAUDEAU, K. NAM, S. GIMENEZ, J.-M. AURY, B. DUVIC, F. HILLIOU, N. DURAND, N. MONTAGNÉ, I. DARBOUX, S. KUWAR, T. CHERTEMPS, D. SIAUSSAT, A. BRETSCHNEIDER, Y. MONÉ, S.-J. AHN, S. HÄNNIGER, A.-S. GOSSELIN GRENET, D. NEUNEMANN, F. MAUMUS, I. LUYTEN, K. LABADIE, W. XU, F. KOUTROUMPA, J.-M. ESCOUBAS, A. LLOPIS, M. MAÏBÈCHE-COISNE, F. SALASC, A. TOMAR, A. R. ANDERSON, S. A. KHAN, P. DUMAS, M. ORSUCCI, J. GUY, C. BELSER, A. ALBERTI, B. NOEL, A. COULOUX, J. MERCIER, S. NIDELET, E. DUBOIS, N.-Y. LIU, I. BOULOGNE, O. MIRABEAU, G. LE GOFF, K. GORDON, J. OAKESHOTT, F. L. CONSOLI, A.-N. VOLKOFF, H. W. FESCEMYER, J. H. MARDEN, D. S. LUTHE, S. HERRERO, D. G. HECKEL, P. WINCKER, G. J. KERGOAT, J. AMSELEM, H. QUESNEVILLE, A. T. GROOT, E. JACQUIN-JOLY, N. NÈGRE, C. LEMAITRE, F. LEGEAI, E. D'ALENÇON, P. FOURNIER. *Two genomes of highly polyphagous lepidopteran pests (Spodoptera frugiperda, Noctuidae) with different host-plant ranges*, in "Scientific Reports", December 2017, vol. 7, n^{o} 1, pp. 1-12 [*DOI :* 10.1038/S41598-017-10461-4], https://hal.inria.fr/hal-01633879

[16] L. LIMA, B. SINAIMERI, G. SACOMOTO, H. LOPEZ-MAESTRE, C. MARCHET, V. MIELE, M.-F. SAGOT, V. LACROIX. *Playing hide and seek with repeats in local and global de novo transcriptome assembly of short RNA-seq reads*, in "Algorithms for Molecular Biology", December 2017, vol. 12, n^{o} 1, 2 p. [*DOI :* 10.1186/S13015-017-0091-2], https://hal.inria.fr/hal-01474524

[17] T. C. MATHERS, Y. CHEN, G. KAITHAKOTTIL, F. LEGEAI, S. T. MUGFORD, P. BAA-PUYOULET, A. BRETAUDEAU, B. CLAVIJO, S. COLELLA, O. COLLIN, T. DALMAY, T. DERRIEN, H. FENG, T. GABALDON, A. JORDAN, I. JULCA, G. J. KETTLES, K. KOWITWANICH, D. LAVENIER, P. LENZI, S. LOPEZ-GOMOLLON, D. LOSKA, D. MAPLESON, F. MAUMUS, S. MOXON, D. R. G. PRICE, A. SUGIO, M. V. MUNSTER, M. UZEST, D. WAITE, G. JANDER, D. TAGU, A. C. C. WILSON, C. VAN OOSTERHOUT,

D. Swarbreck, S. A. Hogenhout. *Rapid transcriptional plasticity of duplicated gene clusters enables a clonally reproducing aphid to colonise diverse plant species*, in "Genome Biology", 2017, vol. 18, n^o 1, 27 p. [*DOI :* 10.1186/s13059-016-1145-3], https://hal-univ-rennes1.archives-ouvertes.fr/hal-01500475

[18] G. Richard, F. Legeai, N. Prunier-Leterme, A. Bretaudeau, D. Tagu, J. Jaquiéry, G. Le Trionnaire. *Dosage compensation and sex-specific epigenetic landscape of the X chromosome in the pea aphid*, in "Epigenetics & Chromatin", June 2017, vol. 10, n^{o} 1, 30 p. [*DOI :* 10.1186/s13072-017-0137-1], https://hal.inria.fr/hal-01555242

[19] A. Rochefort, S. Boukthir, S. Moullec, A. Meygret, Y. Adnani, D. Lavenier, A. Faili, S. Kayal. *Full Sequencing and Genomic Analysis of Three emm75 Group A Streptococcus Strains Recovered in the Course of an Epidemiological Shift in French Brittany*, in "Genome Announcements", 2017, vol. 5, n^{o} 39, e00957 p. [*DOI :* 10.1128/genomeA.00957-17], https://hal-univ-rennes1.archives-ouvertes.fr/hal-01617890

[20] A. Sczyrba, P. Hofmann, P. Belmann, D. Koslicki, S. Janssen, J. Dröge, I. Gregor, S. Majda, J. Fiedler, E. Dahms, A. Bremges, A. Fritz, R. Garrido-Oter, T. S. Jørgensen, N. Shapiro, P. D. Blood, A. Gurevich, Y. Bai, D. Turaev, M. Z. DeMaere, R. Chikhi, N. Nagarajan, C. Quince, L. H. Hansen, S. J. Sørensen, B. K. H. Chia, B. Denis, J. L. Froula, Z. Wang, R. Egan, D. Don Kang, J. J. Cook, C. Deltel, M. Beckstette, C. Lemaitre, P. Peterlongo, G. Rizk, D. Lavenier, Y.-W. Wu, S. W. Singer, C. Jain, M. Strous, H. Klingenberg, P. Meinicke, M. D. Barton, T. Lingner, H.-H. Lin, Y.-C. Liao, G. G. Z. Silva, D. A. Cuevas, R. A. Edwards, S. Saha, V. C. Piro, B. Y. Renard, M. Pop, H.-P. Klenk, M. Göker, N. C. Kyrpides, T. Woyke, J. A. Vorholt, P. Schulze-Lefert, E. M. Rubin, A. E. Darling, T. Rattei, A. C. McHardy. *Critical Assessment of Metagenome Interpretation – a benchmark of computational metagenomics software*, in "Nature Methods", October 2017, vol. 14, n^{o} 11, pp. 1063 - 1071 [*DOI :* 10.1038/nmeth.4458], https://hal.archives-ouvertes.fr/hal-01633525

[21] S. C. Varma, P. Kolin, M. Balakrishnan, D. Lavenier. *Hardware acceleration of de novo genome assembly*, in "International Journal of Embedded Systems", February 2017, vol. 9, n^{o} 1, pp. 74-89 [*DOI :* 10.1504/IJES.2017.10002593], https://hal.inria.fr/hal-01481800

[22] J. A. Wenger, B. J. Cassone, F. Legeai, J. S. Johnston, R. Bansal, A. D. Yates, B. S. Coates, V. A. C. Pavinato, A. Michel. *Whole genome sequence of the soybean aphid, Aphis glycines*, in "Insect Biochemistry and Molecular Biology", January 2017, vol. 18, n^{o} 1, 27 p. [*DOI :* 10.1016/j.ibmb.2017.01.005], https://hal.inria.fr/hal-01555244

[23] V. Wucher, F. Legeai, B. Hedan, G. Rizk, L. Lagoutte, T. Leeb, V. Jagannathan, E. Cadieu, A. David, H. Lohi, S. Cirera, M. Fredholm, N. Botherel, P. A. J. Leegwater, C. Le Beguec, H. Fieten, J. Johnson, J. Alföldi, C. André, K. Lindblad-Toh, C. Hitte, T. Derrien. *FEELnc: a tool for long non-coding RNA annotation and its application to the dog transcriptome*, in "Nucleic Acids Research", 2017, vol. 45, n^{o} 8, 12 p. [*DOI :* 10.1093/nar/gkw1306], https://hal-univ-rennes1.archives-ouvertes.fr/hal-01532061

Invited Conferences

[24] V. Wucher, F. Legeai, L. Bourneuf, T. Derrien, A. Gallot, S. Hudaverdian, S. Jaubert-Possamai, N. Leterme-Prunier, J. Nicolas, H. Seitz, A. Siegel, S. TANGUY, G. Le Trionnaire, D. Tagu. *Integrative genomics and gene networks for studying phenotypic plasticity in the pea aphid*,

in "10th Arthropod Genomics Symposium", Notre Dame, United States, June 2017, https://hal.inria.fr/hal-01566438

International Conferences with Proceedings

[25] 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, https://hal.inria.fr/hal-01666830

[26] S. François, R. Andonov, D. Lavenier, H. Djidjev. *Global Optimization for Scaffolding and Completing Genome Assemblies*, in "International Network Optimization Conference", Lisboa, Portugal, European Network Optimization Group (ENOG), February 2017, 15 p. , https://hal.inria.fr/hal-01499859

[27] 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://arxiv.org/abs/1702.03154 , https://hal.inria.fr/hal-01566246

Conferences without Proceedings

[28] X. Garnier, A. Bretaudeau, O. Filangi, F. Legeai, A. Siegel, O. Dameron. *AskOmics, a web tool to integrate and query biological data using semantic web technologies*, in "JOBIM 2017 - Journées Ouvertes en Biologie, Informatique et Mathématiques", Lille, France, July 2017, 1 p. , https://hal.inria.fr/hal-01577425

[29] X. Garnier, O. Dameron, O. Filangi, F. Legeai, A. Bretaudeau. *Integration of Linked Data into Galaxy using Askomics* , in "Galaxy Community Conference", Montpellier, France, June 2017, https://hal.inria.fr/hal-01576870

[30] C. Guyomar, F. Legeai, C. Mougel, C. Lemaitre, J.-C. Simon. *Multi-scale characterization of symbiont diversity in the pea aphid complex through metagenomic approaches*, in "International Conference on Holobionts", Paris, France, April 2017, https://hal.archives-ouvertes.fr/hal-01638839

[31] A. Meng, E. Corre, P. Peterlongo, C. Marchet, A. Alberti, C. d. Silva, P. Wincker, I. Probert, N. Suzuki, S. Le Crom, L. Bittner, F. Not. *A transcriptomic approach to study marine plankton holobionts*, in "International Conference on Holobionts", Paris, France, April 2017, https://hal.inria.fr/hal-01575069

[32] H.-S. Pham, J.-H. Ferrasse, O. Boutin, N. Alpy, M. Saez. *Analyse thermodynamique du cycle de conversion utilisant le CO2 Supercritique pour une application en réacteur modulaire*, in "SFGP 2017", Nancy, France, July 2017, https://hal.archives-ouvertes.fr/hal-01679087

Scientific Books (or Scientific Book chapters)

[33] G. BENOIT, C. Lemaitre, G. Rizk, E. Drezen, D. Lavenier. *De Novo NGS Data Compression*, in "Algorithms for Next-Generation Sequencing Data", M. Elloumi (editor), Springer, 2017, pp. 91-115 [*DOI :* 10.1007/978-3-319-59826-0_4], https://hal.archives-ouvertes.fr/hal-01633718

Research Reports

[34] L. Bouri, D. Lavenier. *Evaluation of long read error correction software*, Inria Rennes - Bretagne Atlantique ; GenScale, February 2017, n^{o} RR-9028, https://hal.inria.fr/hal-01463694

[35] L. Bouri, D. Lavenier, J.-F. J.-F. Gibrat, V. F. Dominguez Del Angel. *Evaluation of genome assembly software based on long reads*, France Genomique, March 2017 [*DOI :* 10.5281/zenodo.345098], https://hal.inria.fr/hal-01481801

Other Publications

[36] G. BENOIT, P. Peterlongo, M. Mariadassou, E. Drezen, S. Schbath, D. Lavenier, C. Lemaitre. *Simka: large scale de novo comparative metagenomics*, July 2017, 234 p. , JOBIM 2017 - Journées Ouvertes Biologie Informatique Mathématiques, Poster - Acte onlineReference paper: Benoit et al. (2016) Multiple comparative metagenomics using multiset k-mer counting. PeerJComputer Science. https://doi.org/10.7717/peerj-cs.94, https://hal.archives-ouvertes.fr/hal-01595071

[37] C. Benoit-Pilven, C. Marchet, E. Chautard, L. Lima, M.-P. Lambert, G. Sacomoto, A. Rey, C. Bourgeois, D. Auboeuf, V. Lacroix. *Annotation and differential analysis of alternative splicing using de novo assembly of RNAseq data*, November 2017, working paper or preprint [*DOI :* 10.1101/074807], https://hal.archives-ouvertes.fr/hal-01643169

[38] C. Diot, F. Herault, J. Navarro, L. Le Calvez, E. Baeza, C. Klopp, O. Bouchez, D. Esquerre, P. Peterlongo. *Genome specific expression in the liver of mule and hinny duck hybrids*, June 2017, 1 p. , 10. European Symposium on Poultry Genetics (ESPG), Poster, https://hal.archives-ouvertes.fr/hal-01594559

[39] 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*, November 2017, working paper or preprint, https://hal.inria.fr/hal-01634232

[40] C. Guyomar, F. Legeai, C. Mougel, C. Lemaitre, J.-C. Simon. *Multi-scale characterization of symbiont diversity in the pea aphid complex through metagenomic approaches*, July 2017, JOBIM 2017 - Journées Ouvertes en Biologie, Informatique et Mathématiques, Poster, https://hal.archives-ouvertes.fr/hal-01638884

[41] A. Limasset, J.-F. Flot, P. Peterlongo. *Toward perfect reads*, November 2017, https://arxiv.org/abs/1711.03336 - RECOMB 2018 Submission, https://hal.inria.fr/hal-01644163

[42] C. Marchet, L. Lecompte, C. Da Silva, C. Cruaud, J.-M. Aury, J. Nicolas, P. Peterlongo. *De novo Clustering of Gene Expressed Variants in Transcriptomic Long Reads Data Sets*, November 2017, working paper or preprint [*DOI :* 10.1101/170035], https://hal.archives-ouvertes.fr/hal-01643156

[43] C. Marchet, L. Lecompte, A. Limasset, L. Bittner, P. Peterlongo. *A resource-frugal probabilistic dictionary and applications in bioinformatics*, November 2017, https://arxiv.org/abs/1703.00667 - Submitted to Journal of Discrete Algorithms. arXiv admin note: substantial text overlap with arXiv:1605.08319, https://hal.archives-ouvertes.fr/hal-01643162

[44] 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*, November 2017, working paper or preprint [*DOI :* 10.1101/221424], https://hal.archives-ouvertes.fr/hal-01643153

References in notes

[45] 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